Language Identification Using Visual Features

Jacob L Newman, *Member, IEEE,* and Stephen J Cox, *Senior Member, IEEE.*

***Abstract*—Automatic visual language identification (VLID) is the technology of using information derived from the visual appearance and movement of the speech articulators to identify the language being spoken, without the use of any audio information. This technique for language identification (LID) is useful in situations in which conventional audio processing is ineffective (very noisy environments), or impossible (no audio signal is available). Research in this field is also beneficial in the related field of automatic lip-reading. This paper introduces several methods for visual language identification (VLID). They are based upon audio LID techniques, which exploit language phonology and phonotactics to discriminate languages. We show that VLID is possible in a speaker-dependent mode by discriminating different languages spoken by an individual, and we then extend the technique to speaker-independent operation, taking pains to ensure that discrimination is not due to artefacts, either visual (e.g. skin-tone) or audio (e.g. rate of speaking). Although the low accuracy of visual speech recognition currently limits the performance of VLID, we can obtain an error-rate of $< 10\%$ in discriminating between Arabic and English on 19 speakers and using about 30s of visual speech.**

***Index Terms*—Lip reading, speech recognition, language identification, visual speech processing.**

I. Introduction

IT has long been known that visual speech cues can be used by humans to improve speech perception under noisy conditions [1], and the use of visually-derived features to improve automatic speech recognition has been the subject of considerable research [2]–[5]. In the limit, when the audio signal becomes completely inaudible, or is not available, this process becomes lip-reading. Many (primarily deaf) humans can apparently lip-read accurately and fluently, and machine lip-reading techniques have also received considerable attention recently [6]–[8]. A related technology is language identification (LID), which is the technique of identifying automatically the language spoken by a speaker. Audio LID is a mature technology, able to discriminate quite reliably between tens of spoken languages spoken by speakers that are unknown to the system, using just a few seconds of representative speech [9]–[11].

Given the success of LID in the audio domain and the increasing interest in visual speech processing, it is interesting to enquire whether language can be discriminated automatically by purely visual means. Visual language identification (VLID) is an unexplored area of research that is both an interesting research topic in visual speech processing, and a technology that, if successful, would find several useful applications, in, for instance, law-enforcement, and as the first stage of a system that performed visual speech processing. In research terms, it is useful because the task is inherently simpler than lip-reading and it enables us to focus on one of the most difficult aspects of visual speech processing, which is the variation in features across different speakers.

In this paper, we describe initial experiments in the field of VLID. The paper is structured as follows: in Section II, we give relevant background information, including brief reviews of the primary audio LID techniques. Section III describes the datasets we recorded for the task, and section IV describes the techniques and visual features we use. Section V describes our first experiments in speaker-dependent VLID, in which we used bi- and tri-lingual speakers. Section VI extends the techniques to speaker-independent experiments, and includes a description of how we enhanced our features, and an investigation into the effects of skin-tone on discrimination. We end with reflections on what we have learnt and achieved, and some ideas for future work.

II. Background

A. Audio Language Identification

Audio language identification is a mature field of research, with many successful techniques developed to achieve high levels of language discrimination with only a few seconds of test data. The main approaches make use of the phonetic and phonotactic characteristics of languages which are proven to be an identifiable discriminatory feature between languages [12]: see reviews in [13], [9] and [14]. In the next sections, we briefly review the techniques used in these approaches.

1) Phone-Based Tokenisation: There are several approaches to LID which exploit the difference in phonetic content between languages to achieve language discrimination. Such techniques require the training of a phone recogniser, usually comprising a set of hidden Markov models (HMMs), which are used to segment input speech into a sequence of phones.

In an approach called phone recognition followed by language modelling (PRLM) [13], phonotactics is the feature of language used for discrimination. The contention here is that different languages have different rules regarding the syntax of phones, and this can be captured in a language model. In this technique (Figure 1), a single phone recognition system is used to tokenise an utterance using a shared phone set, trained using one language. The phone sequences produced by this system can then be analysed in terms of the co-occurrence (or n-gram) probabilities of phones in an utterance. Statistical models are built using language-specific training data, and these models generate a likelihood score of input utterances being produced by that model. For classification, simple maximum likelihood approaches can be used, or more complex back-end classifiers such as Gaussian mixture models (GMMs), neural networks or support vector machines (SVMs) can be applied. This system can be extended by building PRLM systems using language-specific phone recognisers, and running the recognition systems in parallel (Parallel PRLM =

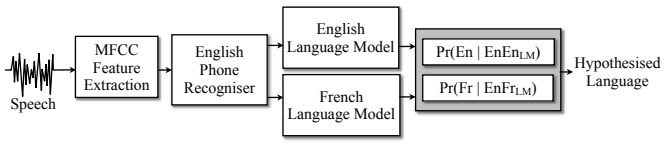

Fig. 1. A system diagram of the phone recognition followed by language modelling approach to audio LID.

PPRLM). A state-of-the-art PRLM system entered into the 2009 NIST language recognition evaluation achieved a mean recognition error of 1.64% on 30-second test utterances in the 23 language closed-set test [10].

2) Gaussian Mixture Model Tokenisation: The tokenisation sub-system within the LID architecture is usually applied at a phone level. [15] presents a variant to the standard PPRLM LID approaches which uses sub-phone, frame-level tokenisation. In this method, a Gaussian mixture model (GMM) is trained for each language from language-specific acoustic data. Each GMM can be considered to be an acoustic dictionary of sounds, with each mixture component modelling a distinct sound from the training data. Given an MFCC frame, the mixture component is found which produces the highest likelihood score, and the index of that component becomes the token for that frame (Figure 2). For a stream of input frames, a stream of component indices will be produced, on which language modelling followed by back-end classification can be performed, as is common in audio LID [15], [16].

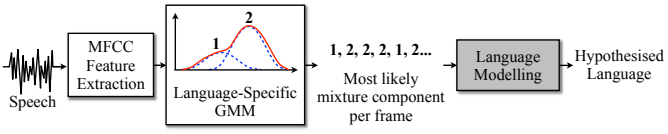

Fig. 2. A system diagram of GMM-tokenisation, instead of phone recognition, as applied to audio LID.

For the NIST 1996 12 language evaluation task [17], [15] report a minimum error rate of 17%, which is higher than standard PRLM techniques. Despite this increase in error rate, several advantages are offered by this approach. Firstly, the training of the tokeniser does not require transcribed data, which simplifies the incorporation of additional languages into the system and is especially advantageous for VLID where there is no agreed protocol for transcriptions. Secondly, there is a reduction in computational cost using this technique compared with phone recognition.

In [18] and [19], a PPRLM LID system similar to that described in Section II-A1 [13] is proposed, except the language models themselves, rather than the scores they produce, become the vectors used by the back-end classifier. Instead of a maximum likelihood or linear discriminant analysis (LDA) back-end, SVMs are built from the bigram language models of the tokenised training data, an SVM for each language, each comparing that target language against all other languages.

B. Human Visual Language Identification Experiments

There has been very little work in investigating whether humans are capable of VLID. [20] reported a number of VLID experiments using human subjects. In one experiment, Spanish-Catalan bilingual subjects were shown 48 silent video clips of people speaking fluently in Spanish or Catalan, two extremely phonologically similar languages. The task was to identify whether the language spoken in the current utterance differed to that of the previous one. Results showed above chance classification accuracy for both Spanish and Catalan dominant participants on this task (57.4% and 60.9% accuracy, respectively), and neither group was found to have a statistically significant advantage over the other. The syllabic count in a sentence was identified as a significant influence on accuracy, with longer test utterances (with higher syllabic counts) resulting in higher recognition accuracies.

C. Visual-Only Speech Recognition

VLID relies on visual speech recognition. Visual speech recognition has generally been studied in the context of audio-visual (AV) speech recognition, and experiments in the field have been mostly speaker-dependent (single speaker) or multi-speaker (the test-set speakers are present in the training-set, although the test data itself is different)—for instance [21], [7], [3]. However, some recent work has focused purely on visual speech recognition, and has extended it to speaker-independent.

[21] and [3] present recognition performances for the visual component of their audio-visual systems. [21] uses both multi-speaker and speaker-independent scenarios. In a digit recognition task using studio recorded video, they obtained 61.47% word accuracy in a speaker-independent task and 76.42% using a multi-speaker task. [7] describes and evaluates two methods of visual feature extraction for integration into an audio-visual speech recogniser. Video-only recognition results are presented for multi-speaker, word-level, isolated letters recognition, using HMMs for speech modelling, and using low resolution grayscale video. The best results presented are 41.9% word accuracy, using active appearance model (AAM) features, and 26.9% using active shape model (ASM) features.

[6] presents results for visual speech-recognition only. They also used letters of the alphabet as test-data but at higher camera resolution and using colour. Their results show that an accuracy of above 80% is achieved for all speakers in a multi-speaker testing scenario, but in speaker-independent tests, the accuracy drops dramatically to below 10%, and in some cases to around chance level. This paper illustrates the strong speaker dependency of the AAM features and cites this as the reason for the poor speaker-independent performance.

A further analysis of the speaker-independent performance of various recognition features is presented in [22]. Using the GRID corpus [23], which consists of speech utterances derived from a highly constrained artificial grammar, and filmed at standard video resolution, [22] concluded that appearance derived features in general out-perform those derived from shape alone, meaning that the appearance of the mouth contains useful information for computer lip-reading. [24] continues this work and suggests some improvements for speaker-independent visual-only speech recognition by using a per speaker z-score normalisation [25], which is used in

this paper (section IV-C), and by applying a hierarchical LDA discriminative training process (HiLDA), which yields a modest improvement to 44% viseme accuracy.

D. Active Appearance Models

Features for visual speech processing are not as well-developed as those for audio. Various feature sets have been tested including DCT coefficients [26], active shape models [27] (ASMs), active appearance models (AAMs) and sieve features [22]. The previous section has described papers in which the recognition performance of some of these features has been studied, and the current finding is that AAM features give the best recognition performance overall [24], despite their poor performance in [6]. AAMs are also routinely used for tracking the contours of the lips and other facial features [28]. For a full exposition of AAMs and how they are used as features in visual speech recognition, see [7].

III. Datasets

Table I presents a brief summary of the content of the two datasets used in the experiments described in this paper. The first dataset, known as United Nations 1 (UN1), was primarily designed for speaker-dependent language recognition using multilingual speakers, and the second, known as United Nations 2 (UN2), was for speaker-independent experiments, specifically for discriminating between English and Arabic speech. Both datasets contain audio and video data of speakers reading the United Nations Declaration of Human Rights. In addition to the audio and video, the datasets contain the tracking information corresponding to the x and y coordinates of a number of contours relating to facial features, principally, the lips.

TABLE I
A brief summary of the UN1 and UN2 datasets.

Dataset	Resolution	FPS	# Languages	# Speakers	# Hours
UN1	576 x 768	25	12	26	≈ 6.5
UN2	1920 x 1080	60	2	35	≈ 7

The UN1 dataset was recorded initially for the speaker-dependent experiments described in Section V, and was also used for initial speaker-independent experiments not reported here. Most of the speakers recorded were competent multilinguals, fluent in either two or three languages. Only a few were truly multilingual, in that they had the same linguistic ability in all languages that they spoke, having learnt to speak them from a very early age [29].

The video camera used for recording the UN1 dataset was a Sony DV domestic video camera. The video format was DV, at a compression ratio of 5:1, which is 25 Mbps ($\approx$ 3.1MBps), the image resolution was 576 x 768 pixels (down-sampled to 480 x 640), and the frame rate was 25 frames per second (progressive scan). We rotated the camera by 90 degrees, so that the dimension with the greatest resolution was the vertical, rather than horizontal dimension. This meant we could frame a speakers's face, whilst occupying most of the frame. Audio was captured from the video camera's built in microphone, and the speech in the audio is intelligible. However, it was not the focus of this database, and therefore no special care was taken to avoid low-level background noise.

The UN2 dataset was recorded for the speaker-independent experiments described in Section VI. We recorded a larger number of speakers than in UN1 for each language, and we recorded only native speakers of a language. The video recorded was also of a higher definition and frame-rate than in UN1: it was captured using a Sanyo Xacti VPC-FH1 domestic video camera, which contains a CMOS sensor. The video was recorded at 60 frames per second (progressive scan), at a resolution of 1920 x 1080 and encoded natively using the MPEG-4 AVC/H.264 codec, at a compression ratio of 1:118.7, which is 24Mbps (3MBps). Although the Sanyo camera enabled us to record at a higher resolution, it was found experimentally that video resolution had little effect on performance. We also captured high quality audio using a tie clip microphone. Word-level transcriptions were manually generated for the English and Arabic speech, and automatically expanded phonetic transcriptions were created (The English transcription contained Arpabet phones and the Arabic transcription a transliteration scheme of the IPA, as described in [30]).

All video data were captured in a studio environment, where lighting was controlled. Each subject sat facing a screen which displayed the text they had to read. They were given a mouse with which to scroll through the document. Subjects were told to sit as still as possible, to face the camera and to avoid occluding their face with their hands. In UN1, the entire head was captured in order to assist the mouth-tracking process, and in case additional facial information was necessary for any further work, and just the mouth region was captured in UN2. Subjects were also advised to carry on reading regardless of any recital mistakes. The text chosen was the United Nations Universal Declaration of Human Rights, because it is freely available in over 300 languages on the web [31]. Using the same text in each language gives some consistency in the style of speaking used and also in the phonetic coverage of the speech. Speakers were required to read about 900 words, which typically took about seven minutes, in each of their fluent languages for UN1, and the entire declaration for UN2 (lasting about 14 minutes). Full details of the recording conditions can be found in [32].

IV. Techniques

A. Building Active Appearance Models

In section II-C we described several computer lip-reading experiments where the recognition features used were derived from an AAM (for appearance), an ASM (for shape) or a combination of both. Although the recognition accuracies reported are far worse than current audio speech recognition systems can achieve, the literature shows that AAMs currently outperform a range of other visual features in these tasks. In our speaker independent experiments we use AAM features. However, in our earlier, speaker dependent experiments, we used ASM features, since they also provide good language discrimination.

To construct an AAM, a selection of training images is marked with a number of points that identify the features of interest on the face. We use the inner and outer lip contours, the jaw line, eyes and eyebrows, totalling around 49 landmark points. The parameters corresponding to non-lip elements are included only for the purpose of assisting tracking capability and are discarded during feature vector generation. The images labelled should represent the extremities in shape and appearance that the model is expected to track, and represent: we label between 10 and 20 images per speaker. The feature points are normalised for pose (translation, rotation and scale), the x and y vectors are concatenated and subject to a PCA, to form the ASM. A representation of part of a shape model is shown in Figure 3. In all cases, where PCA was used to reduce the dimensionality of a set of features, we used the top N PCA components which accounted for 95% of the variance of the data. This means that the dimensionality of the features used varies between different datasets.

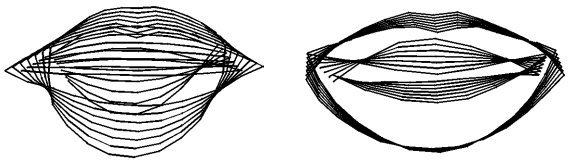

Fig. 3. The first mode (left) and second mode (right) of variation of the shape component of an AAM varying between ± 3 standard deviations from the mean. Lines are separated by one standard deviation. The first mode appears to capture variation due to mouth opening and closing, and the second appears to capture variation due to lip-rounding.

AAM appearance is computed as follows: Each training image is shape normalised by warping it from the labelled feature points, to the mean shape. Our implementation of the AAM uses the RGB colour space. The pixel intensities within the mean shape are concatenated, and the vectors representing each colour channel are then concatenated. The resultant vectors, one for each training image, are subject to a PCA. An example appearance model is shown in Figure 4.

Fig. 4. The mean and first three modes of variation of the appearance component of an AAM.

We use the inverse compositional project-out algorithm [33] to track landmark positions over a sequence of video frames. This algorithm iteratively adjusts the landmark positions on an image by minimising the error between the mean appearance and the appearance contained by the current landmarks, warped to the mean shape. The initial position of the landmarks are set manually, or by their position on the previous frame. If the tracking is inaccurate, it can be necessary to use different initial landmarks, or to adjust the training images so that they better model the variation observed.

To generate the feature vectors used for recognition, all non-lip landmarks are discarded from the training images and the tracked landmarks. New ASM and AAM models are built using the reduced set of training landmarks, and each tracked frame is projected through the models. ASM vectors are the features used in section V for our speaker-dependent work (between 5 and 6 dimensions), whilst concatenated ASM and AAM vectors are used in section VI for our speaker-independent work (8 shape and 48 appearance dimensions). These vectors form a parameter trajectory through the AAM space, one vector for each video frame, corresponding to the words spoken by the speaker being tracked. The eigenvalues for each dimension vary, which means that the scale of each dimension is different. Normalisation is required to prevent individual modes from dominating distance calculations (such as in clustering algorithms) on the basis of their contribution to the overall variation: this is described in section IV-C.

B. Visual Models of Phonemes

Visual speech is typically transcribed into visemes. The usual way of defining visemes is to use expert knowledge to construct a many-to-one mapping from phonemes to visemes, which leads to about 15 visemes compared to about 40 phonemes. This reduction in the number of units reduces the number of possible bigrams of units by about 85% compared with using phone units, and hence reduces VLID performance. Our experience using visemes that were defined by the mapping described in [34] showed that performance was limited, and analysis of this mapping showed that it was highly over-simplified. Although we know that there are several phonemes that cannot be discriminated visually (for instance, it is impossible to detect voicing visually, or place of articulation when this is far back inside the oral cavity), we have found that VLID accuracy is enhanced by training models using video segments corresponding to 42 audio phonemes. We term this representation a "visually-described-phoneme", or VDP. Therefore, in this work, we tokenise our speech in terms of VDPs, rather than visemes.

Tied-State Multiple Mixture Component Triphone HMMs are normally used in state-of-the-art speech recognition systems because of their ability to model coarticulation around a central phone. To build visual triphones (used only in the SI experiments), we manually transcribed the accompanying audio at word level and used a pronunciation dictionary to expand this transcription automatically to phone level. A "flat start" was then applied to the training data so that the segmentation of the AAM frames into VDPs was driven by the data, and not influenced by the audio segmentation. Triphones were built using context-based clustering as described in the HTK manual [35].

C. Enhancing the AAM Features

We examined typical AAM features and found the distribution of values within each dimension to be approximately Gaussian, although means, variances and scale varied from speaker to speaker, and from dimension to dimension. In the experiments on speaker-independent (SI) LID described in section VI, each AAM dimension was z-score normalised per speaker in an attempt to reduce the speaker dependency of

the features. Before normalisation, AAM features are well separated between speakers, meaning that there is no correspondence between the feature vectors for each speaker. After applying the z-score, the relative distance between speakers is much smaller. The normalisation is defined as

$$z_j = \frac{x_j - \bar{x}_j}{\sigma_j} \quad (1)$$

where z_j is the j'th dimension of the normalised vector $\mathbf{z}$, $\bar{x}_j$ is the mean of the speaker's vectors in dimension j, and σ_j is the standard deviation in dimension j.

In these experiments, the data was also linearly interpolated from 60 Hz to 100 Hz, as in [4], in order to provide a suitable number of visual frames to train three state HMMs of VDPs. Although such up-sampling does not, of course, provide any new information, it avoids the problems that are encountered when there are only a few frames per state available.

To improve the discrimination of the features we extract, we also weight the j'th dimension of the feature vector z_j by the mutual information between this feature and the VDP classes. The mutual information was estimated for each dimension by pooling the training vectors and labelling them according to their VDP class. Then, for each dimension of the feature space, the training-data values (over all phone classes) were quantised using a linear quantiser with 16 levels. The mutual information between the classes and z_j is then estimated as follows:

$$I(C, z_j) = \sum_{k=1}^{K} \sum_{l=1}^{L_i} \Pr(C_k, z_j(l)) \log \left(\frac{\Pr(C_k | z_j(l))}{\Pr(C_k)} \right), \quad (2)$$

where $z_j(l)$ is the l'th quantisation level in dimension z_j, $L_j = 16$ and K is the number of classes, e.g. 42 if VDPs are used. By weighting the feature vectors in this way, we give greater weight to the AAM dimensions which are most useful for discriminating the phone classes, whilst giving lower weighting to the least important, which we might expect to be the more speaker-dependent dimensions.

D. Language Modelling and SVM Classification

In the SI experiments, we use a set of ten English speakers to provide training-data: this data is not then used for testing. Because there is not enough Arabic data to partition in this way, only English data is used for visual model training. Bigram language models for the two languages to be discriminated are built from the phone transcriptions of the training-set (generated by the English VDP recogniser). Test data is transcribed into phones, and each language model produces a likelihood for a given utterance, which is length normalised. Back-off weights are calculated and used for unseen bigrams in the test data. Classification is performed using an SVM back-end classifier. For a given utterance in our experiments, two language model likelihoods are produced, as shown in Figure 5. The vector constructed from the likelihoods contains two streams; the first is the ratio of likelihoods of the utterance between the two language models. The second stream applies a linear discriminant analysis (LDA, [36]) transformation to the likelihood scores, builds a Gaussian probability density function (GPDF) from the projected data, and then uses the ratio between the likelihoods from each language. The two streams are then concatenated together. At training time, these vectors are used to build a SVM, which finds the maximum margin hyperplane separating the training data classes. Our SVM uses a Gaussian radial basis function kernel to create a non-linear classifier, as the likelihood scores are not linearly separable. In this task we found that fusion of SVM and LDA outperformed implementations of either LDA or SVMs alone.

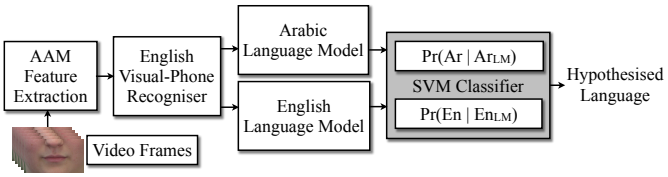

Fig. 5. The system used for speaker independent VLID of English and Arabic.

V. Speaker-Dependent Visual Language Identification

The work in [6] showed that there is strong speaker dependency in the visual features i.e. there is little correspondence between the feature spaces occupied by different speakers. In the preliminary experiments in VLID described in this section, we attempt to discriminate between different languages spoken by the same person. This avoids the complex effects on the features when multiple speakers are used, and hence enables us to focus on the question of whether VLID is even possible. We used the UN1 dataset, described in Section III. The system developed here is based upon a standard audio LID approach, where feature vectors are tokenised and then language models are estimated from the streams of tokens produced by each language.

A. VLID using ASMs and Vector Quantisation

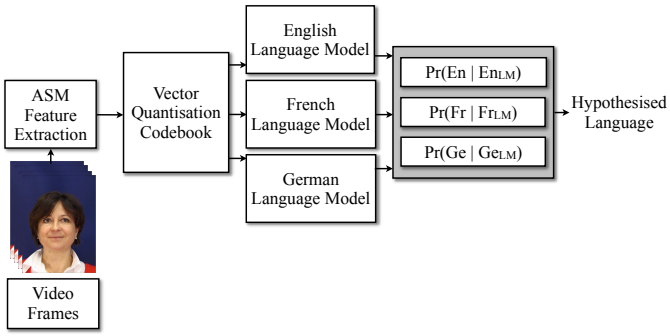

Fig. 6. Unsupervised VLID system diagram, using VQ tokenisation of ASM feature frames.

Figure 6 shows the automatic video language identification system developed here. This system uses vector quantisation (VQ) tokenisation [37] of the feature vectors rather than GMMs. This approach has the advantage that a VQ Bigram "language model" can be constructed for each different language to aid classification.

The video data for each speaker is tracked using an active appearance model (AAM), from which active shape model (ASM) features are derived as described in Section IV-A. The vectors produced by this process are clustered using VQ to produce a VQ codebook. This codebook is then used to tokenise the training data utterances as sequences of VQ symbols. VQ Bigram "language models" for each language recorded by a speaker are then built from these sequences, for each language spoken by that speaker. Unseen bigrams are smoothed to a count of one during generation of the language models.

Test data is transcribed into codewords in the same way as the training data is coded. These codewords are processed by the different language models to produce a likelihood for a given utterance for each of the target languages. Back-off weights are calculated and used for unseen bigrams in the test data. Classification of a test utterance is determined by the bigram language model producing the highest total likelihood for the given utterance. This is calculated by finding the sum of the log probabilities from a language model across all frames in a test utterance, giving the total log probability of a test utterance given a language model.

B. Experiments

Cross-fold validation was used to evaluate the performance of the LID system developed here. An equal number of ASM vectors from each language of a single speaker were divided sequentially and exhaustively to give test utterance durations of 60, 30, 7, 3 and 1 seconds. As an example, if a speaker read the UN Declaration in English (lasting 6 minutes) and French (lasting 7 minutes), the frames in the shorter recital would be divided into 6 one-minute, 12 30-second, 51 7-second, 120 3-second and 360 1-second test utterances. The longer recitals are trimmed to the length of the shorter ones and are partitioned consistently with the shortest one, to ensure balanced training data. A single test utterance is selected from each language and all remaining test data is used for training. Partitioning the data in the way described above means that the number of test utterances for shorter test durations greatly exceeds the number of longer duration utterances, and so a certain difference in error-rate measured on long utterances is less statistically significant than the same difference measured on short utterances. The number of codewords used to vector quantise the data is also an experimental parameter, ranging from 8 to 256 codes. Three speakers were used in these tests: one trilingual speaker (English/French/German) and two bilingual speakers (English/Arabic and English/German). This is a small fraction of the number of speakers available in UN1 database, but the manual work and time involved in preparing the sequences for tracking, and then tracking them, should not be underestimated.

Figure 7 shows the results of tests on an English/French/German trilingual speaker for different numbers of VQ codewords, and shows that performance increases with the duration of the data and the number of codes used. The performance of the three speakers' 256-codeword systems are also shown in Figure 8. This figure shows that the different language combinations tested here are not equally discriminable,

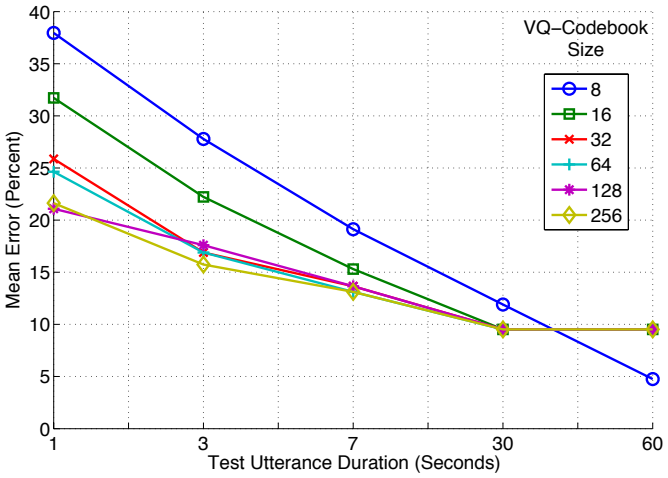

Fig. 7. Results for a VLID system trained on three separate recitals of the UN declaration, read by a single speaker in three different languages; English, French and German. The task is to identify the language from some unseen test data from the same speaker.

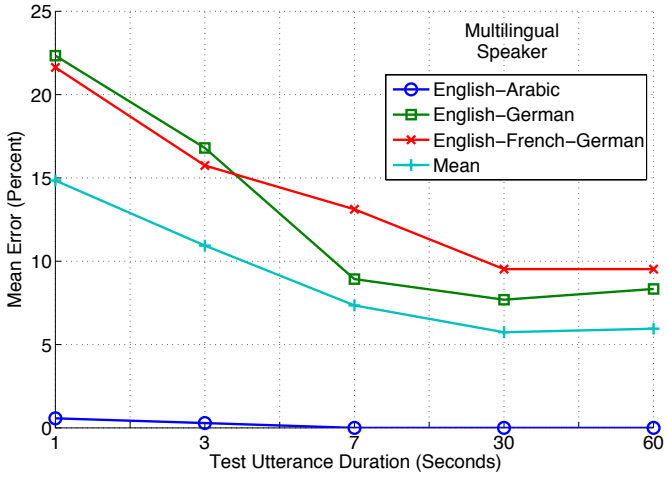

Fig. 8. VLID error for each of the three multilingual speakers tested, and the overall mean. The results shown are for a 256 VQ-codeword system.

either because of true intra-language variation or because of the speaking style of the individual speakers. The figures suggest that classification error decreases with test utterance duration and low error can be achieved for longer utterances. However, it seemed unlikely to us that one second utterances would be sufficient to provide high discrimination performance between three languages. Furthermore, the performance of the eight codeword systems in Figure 7 suggests that eight mouth shapes are sufficient to discriminate between three languages, which was also a surprising finding: audio LID generally uses a symbol set containing over 40 phones, although it is generally used to discriminate many more languages.

C. Bias Due to Speaking Rate and Recording Conditions

We investigated the extent to which unintended effects during recording may have biased results. These might include changes in lighting intensity and colour during the recording, and changes in pose. Since we use only the shape contours

of the mouth, we have largely removed the effect of lighting on our features, although these factors could have affected the performance of the tracker, potentially leading to uneven tracking performance. However, we checked carefully for these effects by examining the spectral distribution of colours in different sessions, and were satisfied that there were no systematic differences between the lighting in each video. Another possible explanation for the low error-rates achieved was that the rate of speech in each recital might be assisting discrimination. When we measured the duration of the recitals for our speakers, we found that they tended to speak their native tongue faster than their other languages i.e. recital speed was correlated with language fluency.

In a low codeword system, each codeword represents a broad area within the feature space, and since rate of speech is linked to rate of change of features, we would expect to see longer runs of the same codeword in slower or less fluent speech. Such a characteristic would be modelled by the bigram language models and would therefore contribute towards classification effectiveness. To test the hypothesis that we were actually measuring differences in rate of speech rather than differences in languages, we performed a similar experiment to the one shown in Figure 7, except that repetitions of the same codeword were ignored and treated as a single occurrence of the codeword. For the eight codeword system, the lowest error-rate in Figure 7, achieved after 60 seconds of data, increased from 5% to 40% after 60 seconds, and for the 16 codeword system, the corresponding rise was from 10% to 27%.

Higher codeword systems were not as affected, as finer clustering of the vector space results in close clusters of data being represented by a number of different codewords, and hence groups of different codewords rather than runs of the same codeword are likely to be observed in slowly-changing speech. The recitals of the speaker that we subjectively judged to have the highest bilingual fluency of the three speakers tested were almost equal in duration. For this speaker, the error-rate rose very little when the experiment described above was run on their utterances.

To determine the sensitivity of our system to variations in speaking rate, we tested it to see whether it could discriminate between three recitations of the same language recorded at different speaking speeds. The system was trained on a single speaker reading three English recitals of the UN declaration in English, read at three different speeds: very slow, a normal reading pace, and very fast. Increasing the rate of speech increases coarticulation, which affects the phonetic content (for instance, assimilation and deletion of phonemes occur more in rapid speech). It is probable therefore that such a large difference in speech rate, as tested here, will alter the phonetic and thus the visual appearance of the speech, resulting in some ability to discriminate between sessions despite containing the same language. Results are shown in Figure 9.

This shows that similar discrimination is achieved to the three language identification task of Figure 7. However, the speed variation we used was extreme: the durations of the readings of the text at fast, medium and slow speeds were 4.6, 6.2 and 7.8 mins respectively, whereas the durations of the texts read in the three different languages by the trilingual speaker were 7.2, 7.8 and 9.0 mins. Hence when different languages were processed, a much smaller speed variation gave about the same discrimination performance, which indicates that rate of speech is not the only effect present.

Finally, we examined whether our system could discriminate between three recording sessions that we had designed to be identical: the same speaker reading the same material in the same language at the same speed. Figure 9 (upper curve) does show a significant reduction in system performance when compared to Figure 7. However, all results shown in Figure 9 are statistically significantly better than the chance error-rate of 66.6%. We can confidently exclude tracking consistency and subtle lighting differences as the reason for this discrimination, since the AAM is trained with equal amounts of data from all sessions and only shape features, rather than shape and appearance, are used for testing. It is more likely that there is a small physical difference between sessions, such as slight pose variations, or that reading performance across sessions was sufficiently different to make the sessions distinguishable.

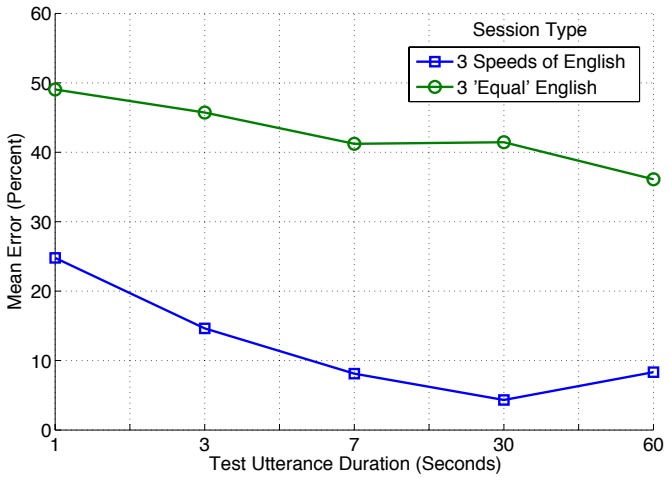

Fig. 9. Showing the mean error-rate across six different codebook sizes (8, 16, 32, 64, 128 and 256 codes) for VLID systems trained on visual speech recorded in two different ways. The upper trace is the error when trained on three separate recitals of the UN declaration in English, spoken by the same speaker, and designed to be identical in rendition. The lower trace is the error when trained on three recitals from the same speaker read at three different speeds: slow, normal and fast.

D. Discussion

Our results for speaker-dependent VLID are somewhat equivocal, because they show that differences in speaking-rate can contribute to language identification. The second experiment showed that apparently even very small differences in a speaker's recital (possibly pose, lighting conditions or recital speed) could be picked up by our system and in some cases were classified with above chance accuracy. However, the fact that different languages spoken at rather similar speeds were as well discriminated as a single language spoken at three extreme speeds indicates that there is a language effect present in these results. Further evidence to confirm that the language

effects presented here are genuine comes from the speaker-independent VLID results reported in the next section.

VI. Speaker-Independent Visual Language Identification

In the previous section, we showed that VLID is possible in speaker-dependent mode by using sub-phonetic units of ASM features, in a manner similar to GMM-tokenisation in audio LID [13]. For speaker-independent VLID, we chose to use the phone recognition followed by language modelling (PRLM) architecture (Figure 1) described in Section II-A, for two reasons: firstly, we cannot continue to use speaker-dependent vector quantisation codebooks in these experiments, and a codebook built using data from many different speakers essentially clusters into areas related to each speaker, and provides little language-specific information. Secondly, experiments in audio LID have shown that phone-based tokenisation outperforms frame-based methods. In these experiments, we use AAM instead of ASM features, because we have shown that they are currently the best features for speaker independent lip-reading, as described in section II-D.

We therefore need to tokenise our visual features using units that are common to all speakers. One such unit is the "viseme", which has been described as the visual appearance of a phoneme in [38], but the exact relationship between phonemes and visemes is unclear and is still a matter for ongoing research [39]. Typically, a single viseme would model many phonetic classes that are considered to be confusable. As described in section IV-B, we found improved language identification accuracy by using a full set of phonemes trained using visual features, which we term "visually-described-phonemes", or VDPs. This raises the difficulty of the strong speaker dependency of our AAM features which was commented upon in section V, and we report in the next section some approaches to ameliorating this dependency. We also use a database that has more speakers (UN2) so that the non language-specific variations mentioned in the previous section tend to be averaged out (although a specific problem with skin-tone is discussed in Section VI-C).

A. Experimental Setup

The task in these experiments is to discriminate between English and Arabic from visual-only information. We selected 19 speakers (10 English and 9 Arabic) from the UN2 dataset for these experiments (Section III). Our testing procedure was 19-fold cross validation, where each of the 19 speakers is held out of the training set in turn, and used for testing instead. As before, for each speaker to be tested, their data was divided sequentially and exhaustively into segments of 1, 3, 7, 30 and 60 seconds.

In our preliminary experiments using our new dataset, we did not build our VDP recogniser from a development set. Instead, we built several English recognisers from our training data, according to which fold of the cross-fold validation we were using (to ensure that the test speaker was not used to train the models). Upon analysis of the likelihood scores from the language models, we consistently saw that unseen data was grouped away from the data used to train the models. We solved this problem by using ten unseen English speakers as a "development set", on which to train our VDP recogniser, which is the approach commonly adopted in audio LID. Without using a development set, we had to generate features for all speakers, for each validation fold, since the feature-generation process for AAMs is data-driven. Using a development set, we can generate a single AAM model from which we can create one set of features for all speakers.

In the language modelling subsystem, counts of phones that did not occur in the training data were smoothed to a count of one. When training the VLID system, the amount of training data for each language was balanced, so that neither language had a bias based upon how well it was modelled. We also length-normalised our language model scores, by using the mean language model score for each utterance. Length normalisation accounts for the fact that longer utterances will have lower likelihoods than shorter ones. Although silence (or its visual equivalent) was recognised in the phone recognition portion of the system, we removed silence from the language modelling subsystem, since silence is not an indicator of the identity of a spoken language.

B. Results

Figure 10 shows the VLID results for the experiments described in this section. Each line is a different speaker. It was found that near perfect language discrimination can be achieved on the corresponding audio data after about 7 seconds of test data. By contrast, most speakers shown here require at least 30 seconds, some 60 seconds, to achieve a similar performance using visual features. There is also greater variation between the performance of speakers in the visual domain when compared to the audio domain, with some speakers only achieving low levels of discrimination with 60s of data. A maximum mean VLID error-rate of 4.64% is achieved after 60 seconds of test data, compared to 0% using audio features. All speakers' results shown in Figure 10, for test utterance durations of 60 seconds, are considered to be above chance ($p < 0.05$).

Table II gives the VDP recognition results for this experiment as % accuracy (as defined in the HTK manual) with a breakdown of the number of correct phones (#C), deletions (#D), substitutions (#S), insertions (#I) and total phones (#T). The test-set accuracy of 17.7% is very low (it compares with a phone accuracy of about 47.6% on the corresponding audio data), but is in line with our recent lip-reading experimentes on other data [22]. Despite this, it is still possible to get some VLID discrimination, given enough data, in this two-language discrimination experiment.

TABLE II
Speaker-independent visually-described-phoneme recognition performance. The results presented are for the English test data used in the experiment described in Section VI-B

Train/Test	Accuracy	# C	# D	# S	# I	# T
Train	31.25	33580	19840	24803	9136	78223
Test	17.71	25364	20015	33063	11474	78442

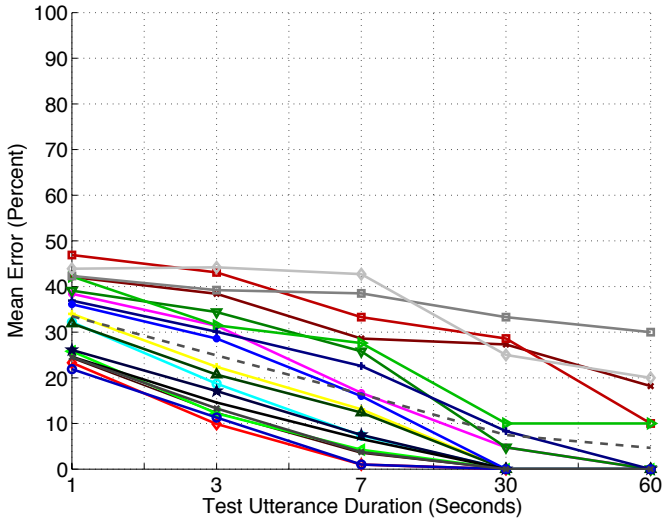

Fig. 10. Speaker-independent VLID results using a PRLM approach with the UN2 visual-only dataset. Each plot line represents the performance of a test speaker. The dashed line is the mean over the speakers.

C. Eliminating the Effect of Skin Tone

The results for VLID accuracy presented in Section VI-B showed that good, two-class, speaker-independent language discrimination could be achieved using visual features. Despite extensive care being taken during the recording process, differences in the recordings could have occurred during data capture, such as focus, colour balance or lighting conditions, and these differences, rather than differences in language, might be responsible for our discrimination. However, the most noticeable difference between the English and Arabic recordings is the skin tone of the speakers: the Arabic speakers all have a darker skin tone than the English speakers. This section reports on experiments designed to answer this question "To what extent is our system's discrimination of English and Arabic based on skin tone rather than genuine language cues?"

Firstly, we note that the skin tone of a speaker is a constant signal over an utterance. Thus, if skin tone were actually the dominant discriminator in our experiments, we would expect that presenting only a few frames to the classifier would be sufficient to achieve good classification accuracy. From Figure 10, we can see that using one second of signal gives a mean error-rate of about 30%. However, we do not know how much of this performance is due to skin tone and how much to language differences. Hence we devised experiments to minimise the effect of colour on our features, by

1) using only shape components (eliminating any appearance information);
2) using a binary representation of the mouth and teeth as colour-free appearance information.

The first experiment removes skin tone altogether from our features. Using shape features also discards any information about the tongue and teeth movements, and any other informative colour variation that is actually separate from skin tone. Therefore, if language discrimination is possible using only shape, we would expect this discrimination to be lower than that obtained using combined shape and appearance features. When the feature dimensions corresponding to appearance are discarded, 8 shape parameters remain out of 56 for the combined shape and appearance features, and these are the dimensions we use for recognition.

Shape parameters are not the optimal skin tone free features that we can extract from the face, as they provide information about only one articulator, the lips. Our second experiment was to use a representation of an additional articulator in order to improve upon our shape-only results. Although we expect the results to be better than using shape-only features, we do not expect performance comparable to that using our original features.

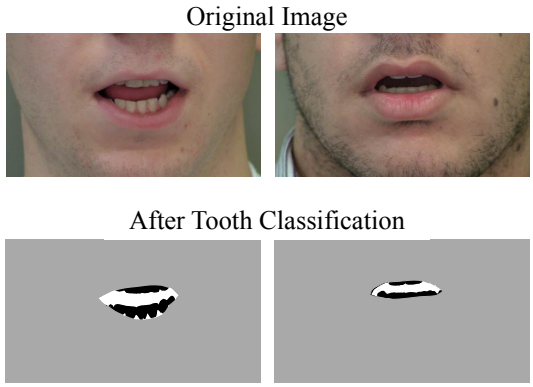

Fig. 11. A binary representation of the teeth. The black pixels have been identified as teeth, and the white pixels as mouth pixels.

Using the hand-labelled images that were created for building the speaker specific AAM trackers, we hand labelled tooth and mouth regions. A 3 x 3 pixel grid was placed over each labelled pixel, then the red, green and blue intensities of those 9 pixels were concatenated and used as the feature vector for each central pixel. A multilayer perceptron classifier was trained speaker-dependently to discriminate between the 27-dimensional vectors corresponding to tooth and non-tooth regions for each speaker. Test vectors were constructed for each pixel contained within the outer lip contour of a test image, and then the recognised regions were set to 255 for non-tooth and 0 for tooth regions, according to the results of the classifier. Finally, small enclosed regions were removed, to reduce the noise caused by a small number of pixels being classified incorrectly.

Using this technique, we were able to represent a sequence of video frames as a series of binary images where the teeth are black and everything else is white (Figure 11). From this, we can perform PCA as in our previous experiments, and run VLID experiments as before. This means that our features contain shape information relating to the lip contours and appearance information corresponding to the position of the teeth within the mouth.

D. Results for skin tone experiments

The remainder of this section presents the results using the shape-only and tooth-pixel recognition features described above in a PRLM system. It should be noted that we cannot present phone recognition accuracy of the Arabic test data

through the English phone recogniser, since we do not have transcriptions of the Arabic speech in terms of English phones.

TABLE III
SPEAKER-INDEPENDENT VISUALLY-DESCRIBED-PHONEME RECOGNITION PERFORMANCE OF THE ENGLISH TEST DATA. THE RESULTS PRESENTED RELATE TO THE AAM SHAPE-ONLY EXPERIMENT DESCRIBED IN SECTION VI-C.

Train/Test	Accuracy	# C	# D	# S	# I	# T
Train	12.58	12476	38933	26914	2536	78223
Test	12.09	11979	36889	29574	2498	78442

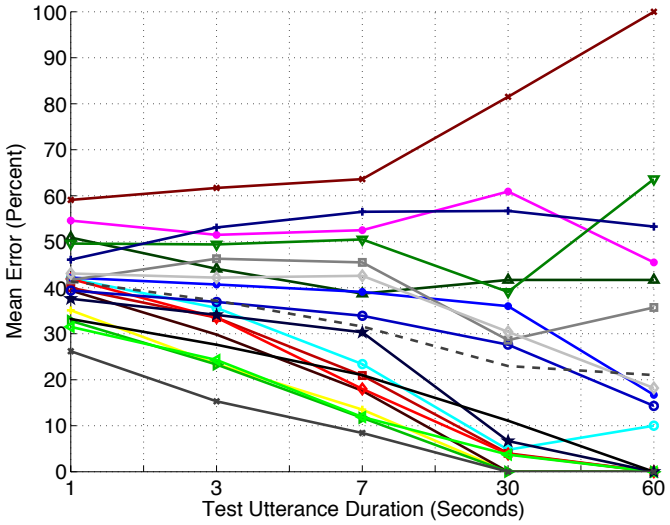

Fig. 12. Speaker-independent VLID results using shape-only recognition features. Each plot line represents the performance of a test speaker. The dashed line is the mean over the speakers.

Table III presents the recogniser performance of the training and test data used to achieve the VLID results shown in Figure 12. Figure 12 shows the VLID results obtained using AAM shape parameters as features for visual-only phone recognition. The results show that whilst around half the speakers can achieve a good identification accuracy with 30 second test utterances, the other half show little or no discrimination. This result is not unexpected, since the amount of articulatory information captured in the shape parameters is limited compared to that derived from the appearance, which is also demonstrated by the lower phone recognition accuracy in Table III compared with Table II. In Figure 12, for the 60 second test utterances, only the results for speakers who achieve lower than 20% mean error after 60 seconds are statistically better than chance. We conclude that some degree of appearance-free language discrimination is possible, although at the cost of overall accuracy.

Figure 13 shows the VLID results using AAM features, generated from tooth segmented video frames, used for visual-only phone recognition. The results show an improvement of performance over the shape-only results in Figure 12, which is expected since we have added the articulatory information of the teeth to our features. A higher recognition accuracy is also present in the phone recognition results shown in Table IV. There are two outlier speakers whose performance

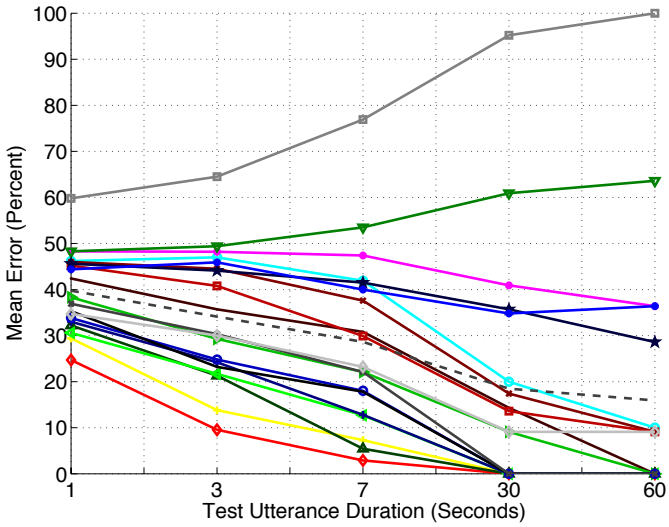

Fig. 13. Speaker-independent VLID results using AAM recognition features derived from tooth-segmented images. Each plot line represents the performance of a test speaker. The dashed line is the mean over the speakers.

actually degrades as test utterance duration increases, and three speakers who achieve only limited language discrimination. These results are lower than those produced by the systems built using full appearance information (Figure 10), which is either due to a loss of visual information regarding modes of variation corresponding to the tongue, and other potentially informative modes, or because there is an effect due to skin tone when full appearance information is used. However, the results do show that reasonable language discrimination can be achieved using colour-free appearance features, which outperform shape alone but are less effective than using full appearance features.

TABLE IV
SPEAKER-INDEPENDENT VDP RECOGNITION PERFORMANCE OF THE ENGLISH TEST DATA. THE RESULTS PRESENTED RELATE TO THE TOOTH-RECOGNITION EXPERIMENT DESCRIBED IN SECTION VI-C.

Train/Test	Accuracy	# C	# D	# S	# I	# T
Train	21.42	25460	22059	30704	8701	78223
Test	15.60	21441	23611	33390	9205	78442

E. Discussion

In this section, we have presented initial results in speaker-independent VLID. We have demonstrated that good speaker-independent language discrimination can be achieved using a PRLM LID architecture, with VDP recognition performed using visual features. We have also shown that some degree of language discrimination is possible using features completely free from appearance or from colour, which shows that the shape and dynamics of the articulators are important for visually-described-phoneme recognition and hence language identification. Although skin tone is the most obvious visual difference between the Arabic and English speakers, there may well be subtler features, such as mouth shape and even speaking style that are characteristic of a language group, and it is important to test whether our system is classifying

these features rather than the spoken language. In [25], we report an ancillary experiment in which we took five bi-lingual speakers and performed speaker-independent language ID on them i.e. each speaker in turn was removed from the training-set and used for testing. Results were about 80% accurate using 60s of speech. This does not prove conclusively that we have eliminated non-language biases, but it does suggest that most of the recognition accuracy is obtained from recognising the language and not other unrelated visual features. The accuracy of our VDP recognisers is very low compared to what can be achieved using audio features. Despite this, we have shown that some degree of speaker-independent visually-described-phoneme recognition is attainable, and that this can be sufficient to achieve good VLID accuracy.

VII. Overall Conclusions and Future Work

In this paper, we have presented an account of initial research into the task of visual language identification (VLID). We have developed two methods for language identification of visual speech, based upon audio LID techniques that use language phonology as a feature of discrimination: an unsupervised approach that tokenises active shape model (ASM) feature vectors using vector quantisation (VQ), and a supervised method of visual triphone modelling using active appearance model (AAM) features. We have demonstrated that VLID is possible in both speaker-dependent and independent cases, and that there is sufficient information presented on the lips to discriminate between two or three languages using these techniques, despite the low phone recognition accuracies that we observed. Throughout, we have taken pains to ensure that the discrimination between languages we have obtained is genuine and not based on differences in the recording or the speakers.

VLID performance is limited by the poor recognition accuracies achieved by our visually-described-phoneme recognisers. Two major issues must be tackled to address this: firstly, the inherent speaker dependence of AAM features, and secondly, the problem of visual ambiguity of phones and therefore the way that they are transcribed. Research into the nature of VDP deletions and the composite units of visual speech [39] may provide a method for dealing with this problem, including the possibility of redefining how visual speech is transcribed, to account for phones that are indiscriminable visually. Many-to-one phone to viseme mappings do not address this problem sufficiently.

Apart from one three-language discrimination task described in section V, this research has focussed on discriminating between two languages. In the future, the number of languages included in the system should be increased to determine how well this approach generalises when the chance of language confusion is higher. Groups of phonetically similar languages could be added to see if they are more confusable than those with differing phonetic characteristics, as well as tonal languages. A range of skin-tones should be present across the languages used, to remove colour as a feature of language. The effect of non-native speech on the visual discrimination of languages could be investigated, as second language speech is shown to affect speech perception in the audio-visual domain [40].

The feature of language that we have used for discrimination is phonology, specifically phonotactics, which governs the allowable sequence of phones in a language. Phonotactics are not the only aspect of language which can be used to differentiate between them. [13] describes the use of phone duration to improve audio LID. In some preliminary work not described here, we tested the technique in [13], where PRLM is performed using double the number of phones, and we saw an increase in identification performance. In this method, each recognised phone is labelled according to whether it is shorter or longer than its mean duration, which doubles the size of a phone set. Another feature of language is rhythm. [41] explains that babies have the ability to distinguish languages based on acoustic rhythm, and [42] suggests that adults also have this ability and furthermore, rhythm is expressed visually. Further work into VLID could therefore focus on incorporating both of these additional language cues and evaluating their contribution to language discrimination.

Acknowledgment

The authors would like to acknowledge the contributions of Dr. Richard Harvey, Dr. Barry Theobald and Dr. Yuxuan Lan to this work. This work was supported by a grant from the UK Engineering and Physical Sciences Research Council, grant number EP/E028047/1.

References

[1] Q. Summerfield, "Lipreading and audio-visual speech perception," *Philosophical Transactions: Biological Sciences*, vol. 335, no. 1273, pp. 71–78, 1992.

[2] G. Potamianos, C. Neti, G. Iyengar, and E. Helmuth, "Large-vocabulary audio-visual speech recognition by machines and humans," in *EUROSPEECH-2001*, 2001, pp. 1027–1030.

[3] L. Liang, X. Liu, Y. Zhao, X. Pi, and A. Nefian, "Speaker independent audio-visual continuous speech recognition," in *IEEE International Conference on Multimedia and Expo (ICME)*, vol. 2, 2002, pp. 25–28.

[4] I. Almajai and B. Milner, "Enhancing audio speech using visual speech features," in *INTERSPEECH-2009*, 2009, pp. 1959–1962.

[5] C. Bregler and Y. Konig, ""Eigenlips" for robust speech recognition," in *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, vol. 2, Apr. 1994, pp. 669–672.

[6] S. Cox, R. Harvey, Y. Lan, J. Newman, and B.-J. Theobald, "The challenge of multispeaker lip-reading," in *International Conference on Auditory-Visual Speech Processing (AVSP)*, 2008, pp. 179–184.

[7] I. Matthews, T. Cootes, J. Bangham, S. Cox, and R. Harvey, "Extraction of visual features for lipreading," *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, vol. 24, no. 2, pp. 198–213, Feb. 2002.

[8] A. Metallinou, C. Busso, S. Lee, and S. Narayanan, "Visual emotion recognition using compact facial representations and viseme information," in *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Mar. 2010, pp. 2474–2477.

[9] M. A. Zissman and K. M. Berkling, "Automatic language identification," *Speech Communication*, vol. 35, no. 1-2, pp. 115–124, 2001.

[10] P. Torres-Carrasquillo, E. Singer, T. Gleason, A. McCree, D. Reynolds, F. Richardson, and D. Sturim, "The MITLL NIST LRE 2009 language recognition system," in *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Mar. 2010, pp. 4994–4997.

[11] NIST, "NIST LRE-2009," http://www.itl.nist.gov/iad/mig/tests/lre/2009/, Jul. 2007, (Last accessed 6/12/10).

[12] P. W. Jusczyk, *The Discovery of Spoken Language*. The MIT Press, 2000.

[13] M. Zissman, "Comparison of four approaches to automatic language identification of telephone speech," *IEEE Transactions on Speech and Audio Processing*, vol. 4, no. 1, pp. 31–44, Jan. 1996.

[14] Y. Muthusamy, E. Barnard, and R. Cole, "Reviewing automatic language identification," *IEEE Signal Processing Magazine*, vol. 11, no. 4, pp. 33–41, Oct. 1994.
[15] P. A. Torres-Carrasquillo, E. Singer, M. A. Kohler, R. J. Greene, D. A. Reynolds, and J. R. Deller Jr., "Approaches to language identification using Gaussian mixture models and shifted delta cepstral features," in *International Conference on Spoken Language Processing (ICSLP)*, 2002, pp. 89–92.
[16] H. Suo, M. Li, P. Lu, and Y. Yan, "Using SVM as back-end classifier for language identification," *EURASIP Journal on Audio, Speech, and Music Processing*, vol. 2008, pp. 1–6, Jan. 2008.
[17] NIST, "NIST LRE-1996," http://www.itl.nist.gov/iad/mig/tests/lre/1996/, Dec. 1995, (Last accessed 6/12/10).
[18] L. Zhai, M. Siu, X. Yang, and H. Gish, "Discriminatively trained language models using support vector machines for language identification," in *IEEE Odyssey: The Speaker and Language Recognition Workshop*, Jun. 2006, pp. 1–6.
[19] H. Li, B. Ma, and C.-H. Lee, "A vector space modeling approach to spoken language identification," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 1, pp. 271–284, Jan. 2007.
[20] S. Soto-Faraco, J. Navarra, W. M. Weikum, A. Vouloumanos, N. Sebastian-Galles, and J. F. Werker, "Discriminating languages by speech-reading," *Perception and Psychophysics*, vol. 69, no. 2, pp. 218–231, 2007.
[21] G. Potamianos, C. Neti, G. Gravier, A. Garg, and A. Senior, "Recent advances in the automatic recognition of audiovisual speech," *Proceedings of the IEEE*, vol. 91, no. 9, pp. 1306–1326, Sep. 2003.
[22] Y. Lan, R. Harvey, B.-J. Theobald, E.-J. Ong, and R. Bowden, "Comparing visual features for lipreading," in *International Conference on Auditory-Visual Speech Processing (AVSP)*, 2009, pp. 102–106.
[23] M. Cooke, J. Barker, S. Cunningham, and X. Shao, "An audio-visual corpus for speech perception and automatic speech recognition," *The Journal of the Acoustical Society of America*, vol. 120, no. 5, pp. 2421–2424, 2006.
[24] Y. Lan, B.-J. Theobald, R. Harvey, E.-J. Ong, and R. Bowden, "Improving visual features for lipreading," in *International Conference on Auditory-Visual Speech Processing (AVSP)*, 2010, pp. 142–147.
[25] J. Newman and S. Cox, "Speaker independent visual-only language identification," in *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Mar. 2010, pp. 5026–5029.
[26] X. Hong, H. Yao, Y. Wan, and R. Chen, "A PCA based visual DCT feature extraction method for lip-reading," in *International Conference on Intelligent Information Hiding and Multimedia Signal Processing (IIH-MSP)*, Dec. 2006, pp. 321–326.
[27] J. Luettin, N. Thacker, and S. Beet, "Speechreading using shape and intensity information," in *Fourth International Conference on Spoken Language Processing (ICSLP)*, vol. 1, Oct. 1996, pp. 58–61.
[28] T. F. Cootes, G. J. Edwards, and C. J. Taylor, "Active appearance models," *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, vol. 23, pp. 681–685, Jun. 2001.
[29] C. Baker, *Foundations of Bilingual Education and Bilingualism*, 4th ed. Multilingual Matters Ltd, 2006.
[30] O. Smrž, "ElixirFM: implementation of functional Arabic morphology," in *Proceedings of the 2007 Workshop on Computational Approaches to Semitic Languages: Common Issues and Resources*. Morristown, NJ, USA: Association for Computational Linguistics, 2007, pp. 1–8.
[31] UN General Assembly, "Universal declaration of human rights," in *General Assembly Resolutions*, vol. 217 A (III), Dec. 1948.
[32] J. Newman, "Language identification using visual features," Ph.D. dissertation, School of Computing Sciences, University of East Anglia, UK, 2011.
[33] I. Matthews and S. Baker, "Active appearance models revisited," *International Journal of Computer Vision*, vol. 60, pp. 135–164, 2004.
[34] S. Lee and D. Yook, "Audio-to-visual conversion using hidden Markov models," in *Proceedings of the 7th Pacific Rim International Conference on Artificial Intelligence: Trends in Artificial Intelligence (PRICAI)*. London, UK: Springer-Verlag, 2002, pp. 563–570.
[35] S. Young, G. Evermann, M. Gales, T. Hain, D. Kershaw, X. A. Liu, G. Moore, J. Odell, D. Ollason, D. Povey, V. Valtchev, and P. Woodland, *The HTK Book (for HTK Version 3.4)*. Cambridge University Engineering Department, 2006.
[36] A. R. Webb, *Statistical Pattern Recognition*, 2nd ed. John Wiley & Sons, Oct. 2002.
[37] Q. Dan, W. Bingxi, and W. Xin, "Language identification using vector quantization," in *6th International Conference on Signal Processing*, vol. 1, August 2002, pp. 492–495.
[38] C. G. Fisher, "Confusions among visually perceived consonants," *Journal of Speech and Hearing Research*, vol. 11, pp. 796–804, 1968.
[39] S. Hilder, B.-J. Theobald, and R. Harvey, "In pursuit of visemes," in *International Conference on Auditory-Visual Speech Processing (AVSP)*, 2010, pp. 154–159.
[40] M. Ortega-Llebaria, A. Faulkner, and V. Hazan, "Auditory-visual L2 speech perception: Effects of visual cues and acoustic-phonetic context for Spanish learners of English," in *International Conference on Auditory-Visual Speech Processing (AVSP)*, 2001, pp. 149–154.
[41] F. Ramus, "Language discrimination by newborns: Teasing apart phonotactic, rhythmic, and intonational cues," *Annual Review of Language Acquisition*, vol. 2, no. 1, pp. 85–115, Oct. 2002.
[42] R. E. Ronquest, S. V. Levi, and D. B. Pisoni, "Language identification from visual-only speech signals," *Attention, Perception, & Psychophysics*, vol. 72, no. 6, pp. 1601–1613, 2010.

Jacob L. Newman received the B.Sc. (Hons.) degree in Computing Science with Electronics and the Ph.D. degree in Language Identification Using Visual Features from University of East Anglia (UEA), Norwich, U.K.

His research interests are audio and visual speech recognition, speech production, and language identification. Currently, he is a senior research associate at the University of East Anglia, and is a member of a team developing a real-time platform for automatic computer lip-reading.

Stephen Cox (M'00–SM'08) received a BSc in physics and music from Reading University and a PhD in speech recognition from the University of East Anglia. He began his career at the UK Government Communications Centre developing signal-processing algorithms. He then joined British Telecoms's research laboratories to lead a team of researchers developing speech recognition algorithms for use on the UK telephone network, and also spent two years at the speech research unit of the UK Royal Signals and Radar Establishment (now Qinetiq). He joined the School of Computing Sciences at UEA as a lecturer in 1991 and was appointed professor in 2003. His research interests include speech recognition, music processing, audio event identification and automatic lip-reading and he is the author and co-author of over 100 publications in these fields.

He was an invited consultant at AT&T Bell Labs, New Jersey in 1994, a visiting scientist at Nuance Communications Inc., CA, in 2000, and an invited researcher at Apple Inc., CA, in 2010. He has acted as a consultant and reviewer for national governments as well as the European Commission, and also consults for industry. He is an ex committee member of the IEEE Speech and Language Technical Committee and was Technical Chair of Interspeech 2009.