AUTOMATIC VISUAL-ONLY LANGUAGE IDENTIFICATION: A PRELIMINARY STUDY

Jacob L Newman and Stephen J Cox

School of Computer Science,
University of East Anglia,
Norwich, UK

ABSTRACT

We describe experiments in visual-only language identification, in which only lip-shape and lip-motion are used to determine the language of a spoken utterance. We focus on the task of discriminating between two or three languages spoken by the same speaker, and we have recorded a suitable database for these experiments. We use a standard audio language identification approach in which the feature vectors are tokenized and then a language model for each language is estimated over a stream of tokens. Although rate of speaking appeared to affect our results, it was found that different languages spoken at rather similar speeds were as well discriminated as a single language spoken at three extreme speeds, indicating that there is a language effect present in our results.

Index Terms— language identification, lip-reading

1. INTRODUCTION

Automatic Language Identification (LID) is a mature technology that can achieve a high identification accuracy from only a few seconds of representative speech [1]. As visual speech processing has developed in the last few years, it is interesting to enquire whether language could be identified purely by visual means. This has practical applications in systems that use either audio-visual speech recognition [2] or pure lip-reading [3] in noisy environments, or in situations where the audio signal is not available. This paper presents a preliminary study in visual LID.

It is known that visual cues can be used by humans in speech processing and contribute to intelligibility [4], but performance in lip-reading is much lower than using audio, even by trained lip-readers. Studies have shown that humans are also capable of identifying language from purely visual cues [5], but again performance is much lower than that obtained using audio signals. The difference between audio and video performance is due to two factors. The first is that speech communication has evolved in such a way that the audio rather than the video-signal has been optimized for error-free communication. The second is that as the video is a secondary communication channel, most people do not develop their lip-reading ability. It is therefore not clear to what extent the task of identifying language from facial features is difficult purely because it is an unusual one that most people are not skilled in, or because the information required for discrimination is not present in the visual features.

The visual communication units of speech are known as visemes [6]. Language identification using only the visual correlates of speech poses a significant challenge as there are fewer distinct visemes than phonemes. Broadly speaking, there is a many to one mapping from phonemes to visemes, increasing the possibility of confusion between speech units, and increasing the difficulty of language identification.

This paper is structured as follows: Section 2 describes the video dataset recorded for this language identification task. The developed visual-only LID system is described in section 3. Section 4 explains the test procedure to be used and presents results produced by the system. Section 5 outlines further work and concludes the paper.

2. APPROACH AND DATABASE

Our previous experiments in lip-reading [3] have shown that the features that we extracted for recognition were highly speaker-dependent. Therefore, we decided that until some features that exhibited greater speaker independence had been developed, our task would be to discriminate between two or more languages read by a single speaker. Hence we chose to record an audio-visual database of multilingual speakers. This approach also has the advantage of focusing on the purely language-specific aspects of the task, and largely eliminating the effect of an individual speaker.

The database recorded contains 21 subjects. These subjects were fluent in at least two different languages, some in three. Typically, these languages consisted of their mother-tongue and a language that they had spoken for several years in an immersive environment. Each subject read a script to a camera in all of the languages in which they were proficient. The subjects were instructed to keep as still as possible, to face the camera, and to avoid occluding their face. They were asked to continue reading regardless of any small mistakes in their recital.

The script chosen was the UN Declaration of Human Rights [7], as translations of this text are available in over 300 languages. Subjects were asked to read up to and including the first 16 articles of the declaration, a text of about 900 words and typically lasting about 7 minutes. The video recorded was 50Hz interlaced scanning at 576x768 pixels, which was changed to 25Hz de-interlaced scanning at 480x640 pixels after post-processing.

3. AAMS AND VECTOR QUANTIZATION

Fig. 1. shows the automatic video language identification system developed here. The video data is tracked using an Active Appearance Model (AAM), as described in section 3.1. The vectors this process produces are first clustered using vector quantization (VQ), detailed in section 3.2, allowing the training data to be tokenized as VQ symbols and bigram language models to be built from the resulting VQ transcriptions. In testing, the AAM vectors are transcribed in the same way and each language model produces a likelihood which is classified by the method outlined in section 3.3.

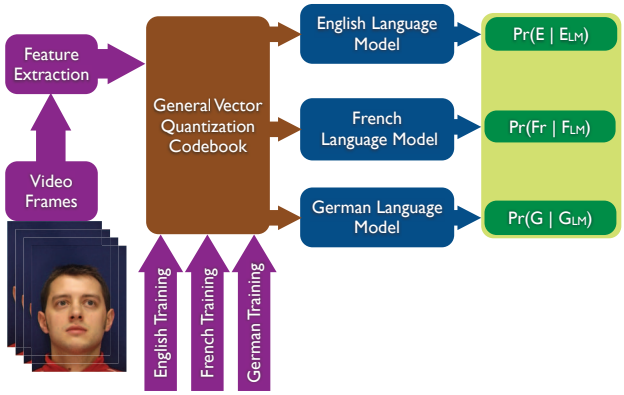

Fig. 1. Visual-only LID System Diagram

3.1. Features: Active appearance models

The AAM tracks the face and lips and produces a vector representing the shape and appearance for each frame of video. However, the parameters corresponding to non-lip elements and also to the mouth appearance are included only to assist tracking capability and are discarded for training and testing, so that the vector consists only of parameters that describe the lip shape. Principal Component Analysis (PCA) is applied to the set of vectors for an individual speaker to reduce the dimensionality. The first few PCA components represent factors such as translation, rotation and scale, and are discarded, leaving about five components to describe lip shape.

AAM generation requires a small number of ground truth frames to build the statistical model used for tracking. The frames selected must represent the extremities in shape that the tracker can expect to encounter. In the system described here, an AAM is built for each speaker using a manual selection of typical frames. These are taken from near the start, middle and end of each language for a single speaker, totaling no more than 15 frames per language.

3.2. Feature Modelling: Vector quantization

Because "visemic" transcriptions are not available for the video data used here (in fact there is no agreed method for transcribing visemes from speech), we use a version of the system described in [8], in which the signal is not transcribed as a sequence of phonemes, or in our case, visemes. In [8], a Gaussian Mixture Model (GMM) was used and the identity of the mixture component that was most likely to have generated a frame was recorded and used as the input to a language model. We have replaced this process with a straightforward vector quantization (VQ) process, which it closely resembles. Once all AAM vectors have been produced from the video data, the designated training segment is vector quantized using a standard k-means algorithm.

3.3. Language model likelihood classification

Bigram language models for each language recorded by a speaker are built from the codeword transcriptions of the training data for each language from that speaker. Unseen codewords 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 and each language model produces a likelihood for a given utterance. 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 probability of a test utterance given a language model. A tie between all language model likelihoods for an utterance is treated as a failure to classify.

4. EXPERIMENTS

Cross-fold validation was used to evaluate the performance of the LID system. Experiments were performed on single speakers to account for the high speaker dependency of the AAM features extracted, as already explained in section 2. An equal number of AAM vectors from each language of a single speaker were divided sequentially and exhaustively to give test utterance durations of either 60, 30, 7, 3 or 1 seconds. As an example, if a speaker read the 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. A single test utterance is selected from each language and all remaining test data is used for training. In each experiment the system must classify a test utterance as one of the two or three languages spoken by this particular speaker. The number of codewords used to vector quantize the data is also an experimental parameter, ranging from 8 to 256.

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, e.g. for the 60 second utterance tests in the three language case in Fig. 4., there are only 21 test utterances in total. Hence, a single mis-classification of a 60 second utterance translates to a 4.8% drop in average percentage accuracy, which although apparently large, may not be statistically significant.

4.1. Initial experiments

The results shown here demonstrate three separate LID experiment results. These include one three language discrimination experiment and two two language experiments. Each figure shows the mean percentage classification accuracy for each duration of test utterance. Tests using codebooks containing between 8 and 256 codewords are presented.

Fig. 2, 3 and 4 show the results of tests on an English/Arabic bilingual speaker, an English/German bilingual speaker and an English/French/German trilingual speaker. They suggest that classification accuracy increases with test utterance duration and high accuracy can be achieved for longer utterances. However, it seemed somewhat improbable to us that one second utterances would be sufficient to provide the high discrimination performance between two languages (as shown in Figure 2) or between three languages (Figure 4). Furthermore, the performance of the eight codeword systems in these figures suggests that eight mouthshapes are sufficient to represent the complete visemic inventory of up to three languages, which was a surprising result. Given these results, it was decided to investigate extent to which unintended effects during recording may have biased results. These would include changes in lighting intensity and colour during the recording and changes in pose. Although we use only the shape contours of the mouth, these factors can affect the performance of the tracker, leading to uneven tracking performance, and

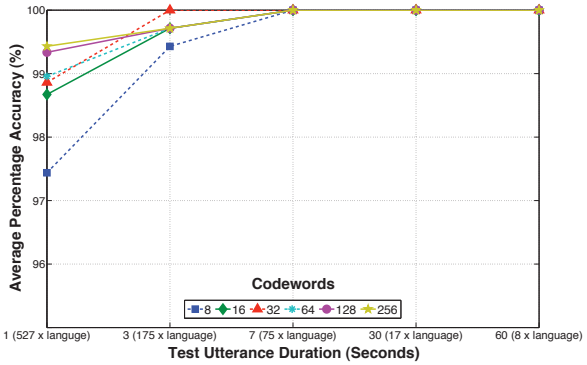

Fig. 2. Testing an English and Arabic Bilingual

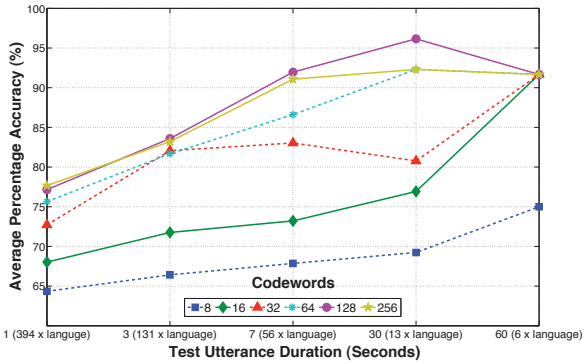

Fig. 3. Testing an English and German Bilingual

this may be reflected in the results. However, we checked carefully for these effects and were satisfied that they were either non-existent or had been satisfactorily removed.

4.2. Removing rate of speech

Another possible explanation for the high accuracy achieved was that the rate of speech might be responsible in part for the high classification accuracy, so that what the system was actually classifying was not the language, but the rate of speaking. Rate of speech is commonly considered to be a measurable characteristic that varies over different languages, but in [9], Roach suggests that this is a simplistic view. However, measurements of the length of the utterances showed that speakers tended to speak their native tongue faster than the other languages.

In a low codeword system, each codeword represents a broad range of 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 modeled 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 Fig. 4., except that repetitions of the same codeword were ignored and treated as a single occurrence of the codeword. The accuracy of the eight codeword system dropped significantly, suggesting that rate of speech was indeed having an effect on performance. Higher codeword systems were not affected, as finer clustering of the vector space results in close clusters of data being represented by a number of different codewords, and hence patterns of different codewords rather than runs of the same codeword are likely to be observed in slowly-changing speech. It is also interesting to observe the performance of the lower codeword systems in Fig. 3., which are produced by the speaker whose bilingual fluency is subjectively judged to be best of those speakers presented in this paper and whose recitals of the UN declaration in each language are almost equal in duration. It would seem rate of speech does indeed effect the classification accuracy, though the extent to which it contributes is not easily determinable from these experiments.

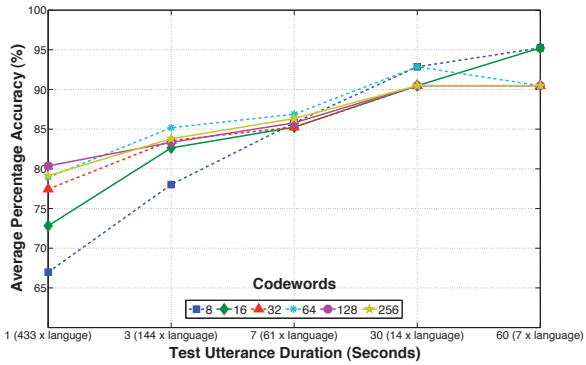

Fig. 4. Testing an English, French and German Trilingual

4.3. Testing rate of speech

As a test of 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. The test here is whether or not each session is identifiable on rate of speech alone. Rate of speech can affect an utterance's phonetic content: for instance, assimilation and deletion of phonemes are more prominent in rapid speech. It is probable therefore that such a large difference in speech rate, as tested here, will alter the phonetic and thus visemic content of the speech, resulting in some ability to discriminate between sessions despite containing the same language.

Fig. 5. shows that similar discrimination is achieved to the three language identification task of Fig. 4. However, the speed variation in Fig. 5. is extreme: the durations of the readings of the text at fast, medium and slow speeds were respectively 4.6, 6.2 and 7.8 mins, whereas the durations of the texts read in three different languages (Fig. 4.) 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 there is an effect of language present.

4.4. Testing session biases

As a final test of the sensitivity of our system to variations in recording conditions, we tested it to see whether it could discriminate be-

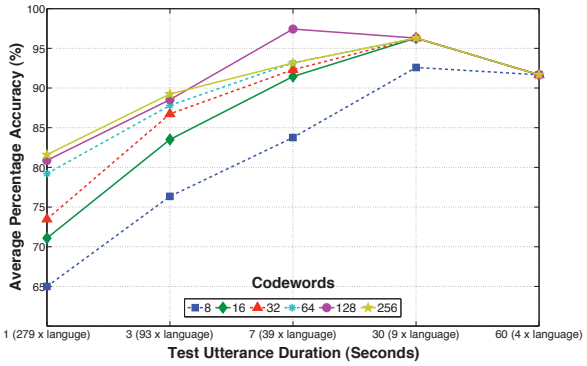

Fig. 5. Testing three speeds of English

tween 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. Without any deliberate variation in rate of speech and in language, the system should be unable to discriminate between sessions and results should therefore be random at around 33%. Fig. 6. does show a significant reduction in system performance when compared to Figs. 4 and 5. However the results are statistically better than random. We can confidently exclude tracking consistency and subtle lighting differences as the causes of this difference, 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.

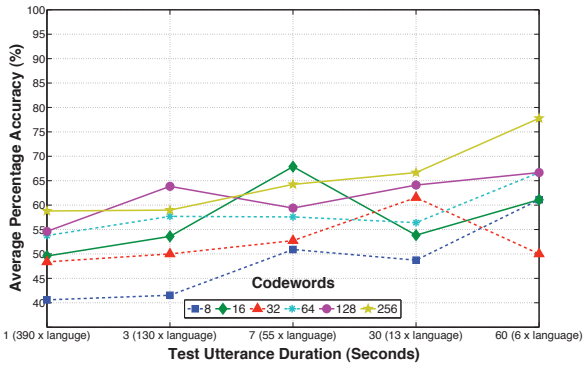

Fig. 6. Testing three 'equal' recitals of English

5. DISCUSSION AND FURTHER WORK

We have presented a preliminary study in identifying language purely from visual features. Because we did not have access to visual features that were independent of the identity of the speaker, we recorded multilingual speakers and attempted to discriminate them reading in two or three different languages. Our results are currently equivocal, because they show that speaking-rate certainly plays a part in identification, and speaking-rate is inextricably bound up with the performance of the speaker in a certain language. A further experiment showed that apparently even very small differences in performance by a speaker were picked up by our system and were classified with above random 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 our results.

To determine the suitability of this technique for visual-only LID, we must first ascertain the contribution to classification performance of visemic content differences caused by language, ignoring non-language variation. The most effective way of achieving this is to average out factors such as the rate of speech, pose and any other potential biases by performing speaker independent language identification on a larger number of speakers. To do this, we will need features that are largely independent of the identity of the speaker. Therefore, future work is to develop such features and test the speaker independence of the this system by testing on a large number of monolingual speakers in a manner similar to audio LID.

6. ACKNOWLEDGMENTS

We would like to acknowledge the contributions of Dr. Richard Harvey, Dr. Barry Theobald and Dr. Yuxuan Lan to this work. This work is supported by UK EPSRC grant number EP/E028047/1.

7. REFERENCES

[1] M. A. Zissman, "Comparison of four approaches to automatic language identification of telephone speech," *IEEE Trans. Speech and Audio Proc*, vol. 4, no. 1, pp. 31–44, 1996.

[2] G. Potamianos, C. Neti, G. Gravier, A. Garg, and A. Senior, "Recent advances in the automatic recognition of audio-visual speech," *Proceedings of the IEEE*, vol. 91, no. 9, pp. 1306–1326, 2003.

[3] S. Cox, R. Harvey, Y. Lan, J. Newman, and B. Theobald, "The challenge of multispeaker lip-reading," *International Conference on Auditory-Visual Speech Processing*, 2008.

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

[5] S. Soto-Faraco, J. Navarra, W. M. Weikum, A. Vouloumanos, N. Sebastián-Gallés, and J. F. Werker., "Discriminating languages by speech-reading," *Perception & Psychophysics*, vol. 69, no. 2, pp. 218–231, 2007.

[6] C. G. Fisher, "Confusions among visually perceived consonants," *Journal of Speech and Hearing Research*, vol. 11, pp. 796–804, 1968.

[7] United Nations, "Universal declaration of human rights," in *General Assembly Resolution*, 1948, vol. 217 A(III).

[8] P. A. Torres-Carrasquillo, E. Singer, M. A. Kohler, R. J. Greene, D. A. Reynolds, and Jr J. R. Deller, "Approaches to language identification using gaussian mixture models and shifted delta cepstral features," *Proceedings of the International Conference on Spoken Language Processing*, pp. 89–92, 2002.

[9] Peter Roach, "Some languages are spoken more quickly than others," in *Language Myths*, L. Bauer and P. Trudgill, Eds., pp. 150–158. Harmondsworth: Penguin, 1998.