

SPEAKER INDEPENDENT VISUAL-ONLY LANGUAGE IDENTIFICATION

Jacob L Newman and Stephen J Cox

School of Computing Sciences,
University of East Anglia,
Norwich, UK

jacob.newman@uea.ac.uk, s.j.cox@uea.ac.uk

ABSTRACT

We describe experiments in visual-only language identification (VLID), in which only lip shape, appearance and motion are used to determine the language of a spoken utterance. In previous work, we had shown that this is possible in speaker-dependent mode, i.e. identifying the language spoken by a multi-lingual speaker. Here, by appropriately modifying techniques that have been successful in audio language identification, we extend the work to discriminating two languages in speaker-independent mode. Our results indicate that even with viseme accuracy as low as about 34%, reasonable discrimination can be obtained. A simulation of degraded accuracy viseme recognition performance indicates that high VLID accuracy should be achievable with viseme recognition errors of the order of 50%.

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 study on speaker independent, visual-only, language identification. In our previous paper [4], we showed that by using sub-phonetic units in a manner similar to GMM-tokenisation in audio LID [1], we were able to discriminate languages spoken by individual speakers. In this paper, we attempt to extend this to speaker-independent discrimination of two languages. This means that we have to abandon the use of the speaker-dependent code-books we used in [4] and use visual units which are common to many speakers.

The visual communication units of speech we use are known as visemes [5]. A viseme is described in [5] as the visual appearance of a phoneme, but the exact relationship of phonemes and visemes in continuous speech is still a matter for ongoing research. Language identification using visemes poses a significant challenge as there are fewer distinct visemes than phonemes. To a first approximation, there is a many to one mapping from phonemes to visemes, so that when visemes are used, there is an increased possibility of confusion between speech units, and hence an increased difficulty of language identification. Also, we have shown the features that we extract from

the face are highly speaker dependent [3], which may limit the performance of a speaker independent system.

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 concludes the paper.

2. APPROACH & DATABASE

Our previous work in visual-only language identification [4] showed clearly that using sub-phonetic units in a manner similar to audio LID systems was sufficient to discriminate languages in speaker dependent experiments. Given that phone recognition generally outperforms sub-phone tokenisation techniques in audio LID, and that preliminary work suggested that using longer temporal information provided greater consistency across speakers, our current work focusses on speaker independent viseme modelling. The approach we have adopted is an adaptation of the parallel phone recognition followed by language modelling approach described in [1]. This change of approach allows us to measure the accuracy of our system in terms of units directly related to speech, which should give a deeper understanding of the speaker independence challenges, and how they can be tackled to facilitate language discrimination.

2.1. VLID Database

For the experiments described here, we used the database described in [4], which consisted of recordings from 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. In this work we have focussed on the task of discriminating between English and French spoken by five speakers, as they are the languages for which we have the most video data. This two class case will allow us to analyse the problems faced more closely and, if successful, we can include more languages later.

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 [6], 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 25 Hz de-interlaced scanning at 480×640 pixels after post-processing.

3. PARALLEL VISEME RECOGNITION FOLLOWED BY LANGUAGE MODELLING

Figure 1 shows the automatic visual language identification system developed for this work. The video data is tracked using an Active Appearance Model (AAM), as described in section 3.1. Audio transcriptions of the video and the AAM vectors are used to train language-specific tied-state viseme HMMs, detailed in section 3.2. Training data is then automatically transcribed as a sequence of visemes, from which language models can be built, and the language model likelihoods are processed using a SVM discriminative classifier, as outlined in section 3.3.

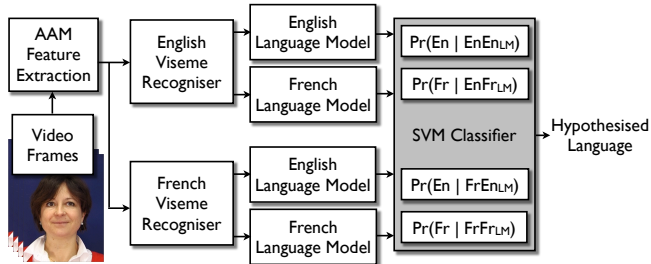


Fig. 1. Visual-only LID System Diagram

3.1. Features: Active Appearance Models

An 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 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 and appearance. 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 between 50 and 60 components to describe combined lip shape and appearance.

AAM generation requires a small number of “ground truth” frames from which a statistical model to be used for tracking is built. 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.

We examined typical AAM features and found the distribution of values within each dimension to be approximately Gaussian, although means and variances varied from speaker to speaker. Given this, each AAM dimension was Z-Score normalised per speaker, per language, in an attempt to reduce the speaker dependency of the features. The Z-Score is a similar normalisation method to Feature Mean Normalisation (FMN), described in [7], except that the distance from the mean is expressed in standard deviations.

The data was also linearly interpolated from 25Hz to 100Hz as in [8] in order to raise the sample rate of the visual signal up to that of typical MFCC data and hence provide a suitable number of visual frames to train three state HMMs. 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 weight the i 'th feature dimension d_i by the mutual information between the feature dimensions and the viseme classes. The mutual information was estimated for each dimension by pooling the training vectors and labelling them according to their corresponding viseme. Then, for each dimension of the feature space, the training-data values (over all viseme classes) were quantized using a linear quantizer with 16 levels. The mutual information between class C_k and d_i is then estimated as follows:

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

where $d_i(l)$ is the l 'th quantisation level in dimension d_i and $L_i = 16$. By weighting the feature vectors in this way, we give greater importance to the AAM dimensions which are most useful for discriminating the viseme classes, whilst giving less weighting to the least important, which we might expect to be the more speaker dependent dimensions.

3.2. Viseme Modelling and Phoneme to Viseme Mapping

Visemes are modelled using tied-state triphone viseme models. Triphone models with a low number of mixture components per model are preferred to monophones with a high number of components, because triphones are better able to model co-articulation. Coarticulation is likely to occur more in visual speech than in audio, as humans are generally not concerned with the clarity of visual articulation. Building viseme HMMs requires labelled training data, and so we map the audio transcriptions of our visual data into visemes.

Viseme mappings define a high-level relationship between phonemes and visemes, and although they do not adequately take account of either the speaker-independent features of visual speech or the effect of coarticulation, they provide a straightforward way of relating the audio and visual domains. Although the mapping between phonemes and visemes is complex, in general, there is a many-to-one relationship from phonemes to visemes. By applying a phoneme to viseme mapping to our audio transcriptions, we effectively represent our visual data as sequences of phoneme super-classes.

There is currently no universally agreed method for transcribing visemes, but several mapping schemes exist. For this work we have chosen the scheme described in [9] to map English phonemes (notated using the ARPabet system) into visemes. A small number of infrequently occurring ARPabet phonemes do not appear in the mapping, and have been manually placed within visually similar viseme groups. Using the mapping in [5], we found where the French IPA pronunciations overlapped with the English Arpa-bet equivalents, and combined the appropriate classes. We merged some French visemes which appeared extremely infrequently and for which we had limited training data. The “short pause” model often used in automatic speech recognition (ASR) was discarded from our set, as we found that the visual articulators do not adopt a consistent position during word transitions, rather they move towards the next spoken phoneme. Table 1 shows the final mapping we applied to the audio transcriptions of our visual data.

3.2.1. Tied-State Multiple Mixture Triphone HMMs

Tied-State Mixture 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 viseme triphones

Table 1. Phoneme to viseme mapping

IPA	Arpabet	Viseme	IPA	Arpabet	Viseme	
p	P	/p/	k	K	/k/	
b	B		g	G		
m	M		n	N		
f	F	/f/	ɸ			
v	V		l	L		
t	T	/t/	ɲ	NG		
d	D		h	HH		
s	S		j	Y		
z	Z		ɪ	IH		
θ	TH		ɪəɾ	IA		
ð	DH		i	IY		
w	W	/w/	ʌ	AH	/ah/	
r	R		ə	AX		
tʃ	CH	/ch/	aɪ	AY	/ao/	
dʒ	JH		ɔ	AO		
ʃ	SH		ɔɪ	OY		
ʒ	ZH		oʊ	OW		
ɛ	EH	/eh/	ʊəɾ	UA		
eɪ	EY		ɑ	AA		
æ	AE		ɑ̃			
aʊ	AW		a			
ɜː	ER		ɒ	OH		
ɛəɾ	EA		o			
ɛ̃			ɥ			
e			y			
ʊ	UH		/uh/	ō		/oo/
u	UW			ø		
œ		/oen/	SIL	SIL	/sil/	
œ̃			SP	SP		-

(or “trisesmes”), we required viseme level transcriptions of our video data. These were generated by manually transcribing the accompanying audio at word level, using HTK to automatically expand the transcription to phone level, and then applying the viseme mapping. A “flat start” was then applied to the training data so that the segmentation of the AAM frames into visemes was data-driven, and not influenced by the audio segmentation.

The BEEP dictionary was used to provide English pronunciations, and we constructed a French pronunciation dictionary manually. Once the audio had been transcribed, we applied the mapping shown in Table 1, to convert our audio transcriptions to visemes. In Table 1, the English-only phonemes are also shown in ARPabet, whilst all phonemes are presented in IPA. From the transcription of our data, we also generated a bigram grammar network for use during recognition.

Using HTK [10] we built three-state, single Gaussian HMMs for each viseme, which were then replicated to form triphone models. Tied-state triphones were then built using hierarchical tree clustering driven by left and right viseme context questions. In a phone recognition system, the most significant rules of coarticulation are known from phonetics knowledge and can be applied to decide which states to examine for automatic clustering. If such rules are not available, as in visual speech, then a data driven approach can be adopted instead to decide which states to tie. Because there are only 16 visemes in our set, as compared to 45 in a typical phoneme set, we can generate all possible combinations of context rules for our visemes, and these provide a computationally manageable number of rules for the clustering process. During clustering, rules that do not satisfy the

state occupancy and likelihood thresholds are ignored, leaving the most appropriate rules for the given parameters. The thresholds we specified retained between 6% and 7% of the total number of states after tying. Finally, the number of mixture components was increased sequentially from one to five, and two was found to give optimal viseme accuracy (presumably for more than two components, the model overfits the training data).

3.3. Language Modelling and SVM Classification

Using the training-set, bigram language models for both languages are built from the viseme transcriptions (recognised from the appropriate language). Test data is transcribed into visemes 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 a SVM back-end classifier. For a given utterance in our experiments, four language model likelihoods are produced, as shown in Figure 1. These are combined into a four-dimensional vector, and at training time, these vectors are used to build a SVM, which finds the maximum margin hyperplane separating the training data classes. The 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 SVMs outperformed LDA-based approaches.

4. EXPERIMENTS

Each of the five subjects we tested was an English/French bilingual speaker. However, only three of them were true bilinguals, having learnt to speak both languages from a very early age.

We use speaker independent, cross-fold validation to evaluate the performance of the LID system: each speaker is held out in turn for testing, and the remaining four are used for training. The time-stamped viseme transcriptions from each language of each single speaker were divided sequentially and exhaustively to give test utterance durations of 60, 30, 7, 3 or 1 seconds. Partitioning the data in this way means that the number of test utterances for shorter test durations greatly exceeds the number of longer duration utterances, and hence improvements on LID accuracies between speakers for longer test utterances may not be statistically significant.

Table 2. Viseme recognition results

ID	%Corr	Acc	ID	%Corr	Acc
1 ENG	50.07	34.33	1 FRE	38.06	29.78
2 ENG	49.95	34.28	2 FRE	47.72	28.80
3 ENG	49.64	34.98	3 FRE	41.24	34.47
4 ENG	49.47	34.69	4 FRE	43.57	33.35
5 ENG	49.64	35.57	5 FRE	44.02	34.70
Mean ENG	49.75	34.77	Mean FRE	42.92	32.22

4.1. Speaker Independent VLID Results

Table 2 shows the speaker independent viseme accuracy of the five speakers used in our experiments. For the values quoted, the grammar scale factor within HTK’s Viterbi recognition tool was set to 0.3 and the word insertion penalty to -20 . These were optimum values (for recognition accuracy, not VLID), determined empirically.

Figure 2 shows the VLID results for our system. The lowest mean error of 12.2% is achieved with 60 seconds of test data. It is

reassuring to note that the VLID error-rate decreases monotonically with utterance length for all speakers except for speaker number 3, whose error rate remains about chance. However, we noted that audio LID was significantly worse for this speaker, and examination of their video recordings showed that they exhibited some unusual visual speech, in particular very prominent top teeth, and a minor speech defect. These factors probably contributed to poor performance, but it is these kind of issues that we wish to examine with a larger database of speakers.

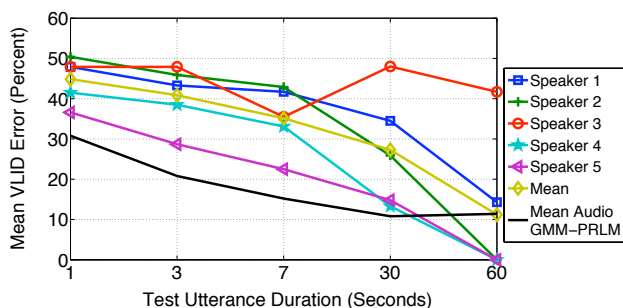


Fig. 2. Speaker independent VLID results

4.2. Simulated Viseme Error on VLID Performance

The accuracy of the viseme recogniser used in our VLID experiments is very low compared to recognisers in audio phone recognition systems. Given that our VLID accuracy appears to be limited by our viseme recognisers, it is interesting to investigate at what level of viseme accuracy high VLID recognition is attainable. Furthermore, we would like to know how degradation of viseme accuracy affects VLID performance. We tested this by simulating different viseme recognition accuracies for our 5 speakers, with speaker 1 used as test data. We begin by using the ground-truth (oracle) viseme transcriptions of the English and French texts to construct bigram language models. Then, for each speaker, we reduce the transcription accuracy by using a model of the pattern of confusions made by an individual speaker. This model enables us to insert substitutions, deletions and insertions into the transcription at effectively any “error-rate” we choose. We then perform VLID using test data from one of the speakers.

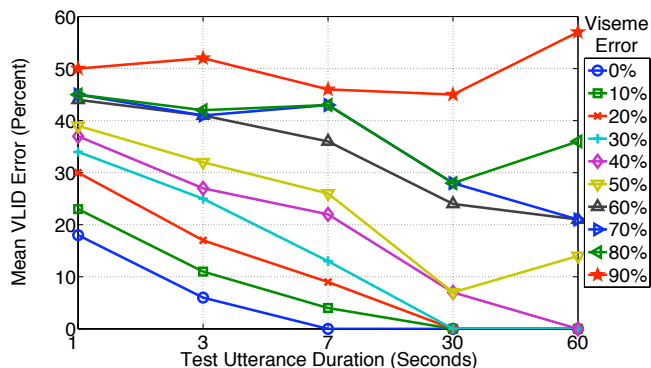


Fig. 3. The effect of viseme accuracy on VLID recognition

The results of this simulation are shown in Figure 3. They indicate that even with an error-rate around 40%, 100% VLID accuracy is achievable with 60 seconds of test data.

5. DISCUSSION AND FURTHER WORK

This preliminary work on a two language discrimination problem using only five speakers indicates that speaker independent visual language discrimination is possible, but it is limited by the low accuracy currently obtainable by viseme recognisers, as suggested by the simulation experiment described in Section 4.2. Also, English and French are both Indo-European languages, sharing many of the same phonetic and phonological characteristics. Therefore, it is possible that more dissimilar languages may be discriminated more easily. Better discrimination may also be achieved by training on a larger set of speakers, and crucially, by developing different visual features that are less speaker-dependent. Further work will concentrate on developing improved visual features and recording a larger database of subjects speaking in more diverse languages.

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] J. L. Newman and S. J. Cox, “Automatic visual-only language identification: A preliminary study,” *2009 IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 4345–4348, 2009.
- [5] C. G. Fisher, “Confusions among visually perceived consonants,” *Journal of Speech and Hearing Research*, vol. 11, pp. 796–804, 1968.
- [6] United Nations, “Universal declaration of human rights,” in *General Assembly Resolution*, 1948, vol. 217 A(III).
- [7] G. Potamianos and A. Potamianos, “Speaker adaptation for audio-visual speech recognition,” *Proc. EUROASPEECH*, pp. 1291–1294, 1999.
- [8] I. Almajai and B. Milner, “Maximising audio-visual speech correlation,” in *Proc. AVSP*, 2007.
- [9] S. Lee and D. Yook, “Audio-to-visual conversion using hidden markov models,” *PRICAI '02: Proceedings of the 7th Pacific Rim International Conference on Artificial Intelligence*, pp. 563–570, 2002.
- [10] Steve Young et al., *The HTK Book (for HTK Version 3.4)*, Cambridge University Engineering Department, 3.4 edition, 2006.