An Evaluation of Visual Speech Features for the Tasks of Speech and Speaker Recognition

Simon Lucey

Advanced Multimedia Processing Laboratory
Department of Electrical and Computer Engineering
Carnegie Mellon University
Pittsburgh PA 15213, USA
slucey@ieee.org

Abstract. In this paper an evaluation of visual speech features is performed specifically for the tasks of speech and speaker recognition. Unlike acoustic speech processing, we demonstrate that the features employed for effective speech and speaker recognition are quite different to one another in the visual modality. Area based features (i.e. raw pixels) rather than contour features (i.e. an atomized parametric representation of the mouth, e.g. outer and inner labial contour, tongue, teeth, etc.) are investigated due to their robustness and stability. For the task of speech reading we demonstrate empirically that a large proportion of word unit class distinction stems from the temporal rather than static nature of the visual speech signal. Conversely, for the task of speaker recognition static representations suffice for effective performance although modelling the temporal nature of the signal does improve performance. Additionally, we hypothesize that traditional hidden Markov model (HMM) classifiers *may*, due to their assumption of intra-state observation independence and stationarity, not be the best paradigm to use for modelling visual speech for the purposes of speech recognition. Results and discussion are presented on the M2VTS database for the tasks of isolated digit, speech and text-dependent speaker recognition.

1 Introduction

It is largely agreed upon that the majority of visual speech information stems from a subject's mouth [1]. The field of audio-visual speech processing (AVSP) is still in a state of relative infancy, during the period of its short existence a majority of the work performed has been towards the goal of finding the best mouth representation for the tasks of audio-visual speech and speaker recognition. Usually, these representations are based on the techniques used to initially locate and track the mouth, due to their ability to parametrically describe the mouth in a compact enough form for use in statistical classification.

This paper concentrates on the evaluation of area features as opposed to contour features, due to their robustness and stability. Area based representations are concerned with transforming the whole input region of interest (ROI) mouth intensity image into a meaningful feature vector. Contour based representations are concerned with parametrically atomizing the mouth, based on a priori knowledge of the components of the mouth (i.e. outer and inner labial contour, tongue, teeth, etc.). In a recent paper by Potamianos et al. [2] a review was conducted between area and contour features for the tasks of speechreading on a large audio visual database. In this paper it was shown that area representations obtained superior performance. Area based representations of the mouth were shown to be robust to noise and compression artifacts and are the mouth representation of choice in current AVSP work.

It is widely accepted that for acoustic speech and speaker recognition applications cepstral [3] features work well in *both* applications respectively. Like many aspects of acoustic speech processing, this rationale has been applied to visual speech processing applications with minimal analysis and evaluation of the validity of such an assumption in the visual modality. In this paper we explore a number of visual speech representations for the tasks of speech and speaker recognition and demonstrate that the modelling of visual speech for the tasks of speech and speaker recognition *are* different in terms of the features and classifiers used.

2 A brief review of area based representations

The most common technique used to gain a holistic compact representation of a mouth is through the use of principal component analysis (PCA) [4], which attempts to find a subspace the main linear modes of variation, on the mouth ROI intensity image. Linear discriminant analysis (LDA) [5] generates a subspace based on a measure of class discrimination. LDA representations have become extremely useful in AV speech [6, 5] and speaker recognition applications. PCA and LDA are referred to as data driven, as they both require training observations of mouth ROI images to create their compact representation of the mouth. Other data-driven transforms have been employed on the mouth region, such as maximum likelihood linear transform (MLLT) [6] and independent component analysis (ICA) [7], albeit with minimal improvement to traditional PCA and LDA techniques.

Non-data driven transforms have been previously used such as the discrete wavelet transform (DWT) [2], discrete cosine transform (DCT) [6] or multiscale spatial analysis (MSA) [8] directly or as pre-processing stage for visual feature extraction. These non-data driven approaches have the benefit of not being dependent on a training ensemble, but bring minimal a priori knowledge about the mouth to the problem of visual speech and speaker recognition.

3 Evaluation of Speech Features

The actual evaluation of visual speech features is not an easy task as an inherent problem with extracting speech features is in getting an accurate measure of how well a given speech feature works when compared against another. Generally an accurate measure of the quality of visual features is indicative of how well it performs in the task it is being used for, which in this case is visual speech and text dependent speaker recognition. As previously mentioned, only area features shall be investigated in this paper due to their robustness and ability to holistically represent the mouth. Data-driven feature extraction approaches were investigated solely in this evaluation due to their natural ability to bring a priori knowledge of the mouth to the representation. For purposes of notation the mouth image matrix $\mathbf{I}(x, y)$ is expressed as the vectorized column vector $\mathbf{y} = vec(\mathbf{I})$. The tasks of speech and speaker recognition were tested with the following visual features,

PCA: in which PCA was used to create a twenty dimensional subspace $\boldsymbol{\Phi}_{PCA}$ preserving the 20 highest linear modes of mouth variation. This feature extraction approach was employed for both speech and speaker recognition.

SLDA: in which LDA was used to create a twenty dimensional subspace $\boldsymbol{\Phi}_{SLDA}$ for the speaker recognition task using a priori knowledge of the subject classes to generate the 20 most discriminant basis vectors.

MRPCA: in which the mean removed mouth sub-image $\mathbf{y}^*$ is calculated from a given temporal mouth sub-image sequence $\mathbf{Y} = \{\mathbf{y}_1, \ldots, \mathbf{y}_T\}$ such that,

$$\mathbf{y}_t^* = \mathbf{y}_t - \overline{\mathbf{y}}, \quad \text{where } \overline{\mathbf{y}} = \frac{1}{T}\sum_{t=1}^{T} \mathbf{y}_t \tag{1}$$

This approach is very similar to cepstral mean substraction [3] used on acoustic cepstral features to improve recognition performance by providing some invariance to unwanted variations. In the visual scenario this unwanted variation usually stems from subject appearance. Mean-removal PCA (MRPCA) uses these newly adjusted $\mathbf{y}^*$ mouth sub-images to create a new twenty dimensional subspace $\boldsymbol{\Phi}_{MRPCA}$ preserving the 20 highest modes of mean removed mouth variation. This approach was first proposed by Potamianos et al. [2] for improved visual speech recognition performance.

WLDA: in which LDA was used to create a nine dimensional subspace $\boldsymbol{\Phi}_{WLDA}$ for the speech recognition task using a priori knowledge of the word classes

to generate the 9 most discriminant basis vectors. Mean removal, similar to the approach used for MRPCA, was first employed to remove unwanted subject variances from the WLDA feature extraction process.

A compact representation of the mouth sub-image $\mathbf{y}$ can be obtained by the linear transform,

$$\mathbf{o} = \boldsymbol{\Phi}'\mathbf{y} \tag{2}$$

such that $\mathbf{o}$ is the compactly represented visual speech observation feature vector. Illumination invariance was obtained by normalising the vectorised mouth intensity sub-image $\mathbf{y}$ to a zero-mean unit-norm vector. For the generation of the LDA subspaces, PCA was first employed to preserve the first 50 linear modes of variation, in order to remove any low energy noise that may corrupt classification performance. For all subspaces, shots one to three of the M2VTS [9] database were used as training mouth observations, with shot four being used for testing in the speech and speaker recognition tasks. In all cases delta (i.e. first order derivative) features were appended to static features.

3.1 Training of hidden Markov models

Hidden Markov models (HMMs) were used to model the video utterances using HTK ver 2.2. [3]. The first three shots of the M2VTS database were used to train the visual HMMs with shot four being used for testing. The database consisted of 36 subjects (male and female) speaking four repetitions (shots) of ten French digits from *zero* to *nine*. In the task of speech recognition the word error rate (WER) was used as a measure of performance for the ten digits being recognized in the M2VTS database.

Speaker recognition encapsulates two tasks, namely speaker identification and verification. Speaker error rate (SER) was used to gauge the effectiveness of visual features for speaker identification. The SER metric was deemed useful enough for gauging the effectiveness of visual features in speaker recognition as good performance in the speaker identification task generally translates well for the verification task. Due to the relatively small size of the M2VTS database and the requirement for separate speaker dependent digit HMMs, all speaker dependent HMM digit models were trained by initializing training with the previously found speaker independent or *background* digit model. This approach prevented variances in each model becoming too small and allows each model to converge to sensible values for the task of text dependent speaker recognition.

3.2 Speech recognition performance

Table 1 shows the WER for the task of digit recognition on the M2VTS database. Raw PCA features have the worse WER performance out of all the visual features evaluated. There is little difference between the MRPCA and WLDA area representation of the mouth in terms of WER at the normal video sample rate

of 40ms, with WLDA visual features performing slightly better. Acoustic MFCC features were also evaluated in Table 1 for comparison with its visual counterparts. The train and test sets of each feature type were evaluated in terms of WER. The difference between train and test WERs is very important as this gives an indication of how undertrained a specific speech recognition classifier is using a certain type of feature [10]. The train WER is also very important as it gives a rough estimate of the lower Bayes error for that feature representation, with the test WER giving an estimate of the upper Bayes error. Both train and test errors are essential to properly evaluate a feature set.

There are very large differences between train and test WERs for all visual feature sets in comparison to the differences seen in the acoustic MFCC feature set. Additionally, the test WERs for all visual features are quite large, which is in stark contrast to the acoustic MFCCs which received negligible error. This may indicate the inherent variability of the chosen visual features is higher than those found in conventional acoustic features, or that the visual features do not provide enough distinction between word classes using a standard HMM classifier. Similar results were received by Cox et al. [10] pertaining to the undertrained nature of standard HMM based visual speech recognition classifiers.

Initially, one may assume the undertrained nature of the visual HMM classifiers may be attributed to the acoustic modality having four times as many training observations as the visual modality. This is due to the acoustic speech signal being sampled at a 10m intervals, with the visual speech signal being sampled at a coarser 40ms interval. To partially remedy this situation, the visual features were up-sampled[1] to 10ms intervals using simple linear interpolation. Inspecting Table 1 one can see that the WER increases when testing is performed on the interpolated visual features using the same topology (i.e. number of states and mixtures) HMM classifier for all visual feature types. However, when the number of HMM states is increased the WER performance of all interpolated visual features improves. For PCA and MRPCA representations the WER actually surpasses those seen at normal sample rates. The interpolated MRPCA based HMM classifier with extra states receives an WER that marginally surpasses that for the normally sampled WLDA classifier. Additionally, the train WER for the interpolated MRPCA classifier, with extra states, is half of that for the normally sampled WLDA classifier, indicating that the increase in classifier complexity may provide additional word class distinction. The interpolated WLDA features, using an increased number of states, still receives a poorer WER than realised with the originally sampled WLDA features with less states. The lack of performance improvement in the WLDA representations, using interpolation with an increased number of states, indicates that some vital discriminative information pertaining to the temporal nature of the utterance is being thrown away in comparison to the PCA and MRPCA representations. This could be

[1] Interpolation of visual features occurred prior to the calculation of delta features, which were used in all experiments. It must be noted that when interpolation was employed on static and previously calculated delta visual features minimal change in WER was experienced.

Features	(Dim)	Sampling	HMM Topology		WER(%)	
			Mixtures	States	Train set	Test set
PCA	40	40ms	3	3	14.19	31.43
PCA	40	10ms	3	3	21.43	39.71
PCA	40	10ms	3	9	8.07	28.57
MRPCA	40	40ms	3	3	9.71	25.71
MRPCA	40	10ms	3	3	13.52	30.57
MRPCA	40	10ms	3	9	5.33	23.14
WLDA	18	40ms	3	3	10.38	23.43
WLDA	18	10ms	3	3	17.11	33.43
WLDA	18	10ms	3	8	12.76	28.57
MFCC	26	10ms	3	3	1.44	1.62

Table 1. WER rates for train and test sets on the M2VTS database (note best performing visual features have been highlighted).

attributed to the majority of discriminatory information between words being contained in the temporal nature of the pronunciation *not* the static appearance. A major drawback in WLDA feature extraction seems to stem from its inability to form a discriminative subspace based on the dynamic, not just static, nature of the signal. Potamianos et al. [5] devised an approach to circumvent this limitation by incorporating contextual information about adjacent frames into the construction of a discriminative subspace. Although showing some improvement, this approach fails to address some of the fundamental problems associated with using a standard HMM classifier for speech reading.

The performance improvement from the interpolation of PCA and MRPCA features along with the increase in HMM states for their respective HMM classifiers can be considered to be counter intuitive, as no extra information is being added to the interpolated visual features apart from the delta features which are dependent on the sample rate of the signal. The benefit of interpolating visual features can be understood from work done by Deng [11] concerning standard HMM based speech recognition. Deng has argued that the use of many states in a standard HMM can approximate continuously varying, non-stationary, patterns in a piecewise constant fashion. Further, it was found in previous acoustic speech recognition work [11], that as many as ten states are needed to model strongly dynamic speech segments in order to achieve a reasonable recognition performance. Similar results were found by Matthews et al. [8] for visual speech recognition where as many as nine states were required, after visual feature interpolation, to achieve reasonable WERs.

It has been postulated by Deng [11] that employing extra states in a standard HMM to better model the non-stationary dynamic nature of a signal in a piece-wise manner has obvious shortcomings. This is due to the many free and largely independent parameters needing to be found by the addition of extra states which requires a large amount of training observations for reliable classification. The problems concerning the lack of training observations can be partially combated through the interpolation. Such trends can however, be much

more effectively and accurately described by simple deterministic functions of time which require a very small number of parameters, as opposed to using many HMM states to approximate them piecewise constantly. This indicates that, unlike the acoustic modality, the use of a standard HMM may be suboptimal for the purposes of modelling the non-stationary nature of the visual speech modality effectively for speech recognition.

3.3 Speaker recognition performance

Table 2 shows the SER for the task of text dependent speaker identification. The use of SLDA in this instance is of considerable benefit over the traditional PCA representation of the mouth. Intuitively, this makes considerable sense as a person's identity can be largely represented by the static representation of that person's mouth. This result differs to those found in visual speech recognition, which found the discriminant nature of WLDA to be of limited use due to the majority of the class distinction between words existing in the temporal correlations in an utterance rather than the static appearance of the mouth. The

Features	(Dim)	Sampling	HMM Topology		SER(%)	
			Mixtures	States	Train set	Test set
PCA	40	40ms	2	2	0.38	28.00
PCA	40	10ms	2	2	0.67	28.29
SLDA	40	10ms	2	2	0.19	19.71
SLDA	40	40ms	2	2	0.19	19.71
MFCC	26	10ms	3	2	0.00	9.72

Table 2. SER for train and test sets on the M2VTS database (note best performing visual features have been highlighted).

up-sampling of visual features was also investigated, but from an exhaustive search through HMM topologies, there was no improvement in SER from the optimal topologies used at the normally sampled rates. This result can be attributed to two things. Firstly, there is an inherent lack of training observations for generating a subject dependent digit HMM, making the generation of suitably complex HMMs difficult. Secondly, the piece-wise temporal approximation made by a standard HMM suffices for the task of visual speaker recognition due to its natural ability to discriminate based on static features, as indicated by the superior performance of SLDA over PCA features. Interestingly, the performance of the acoustic and visual classifiers are relatively close, with both classifiers being marginally undertrained. This result was to be expected due to the lack of training data associated with each subject and digit.

4 Discussion

In this paper feature extraction techniques for the visual speech modalities, pertaining to the tasks of speech and speaker recognition, were evaluated. For speechreading it was shown that MRPCA mouth features, at an interpolated sample rate, gave superior WERs over all those evaluated. Although, WLDA features, based on a static discriminant space, perform almost as well and do not require interpolation and have a much smaller dimensionality. For both feature sets the benefit of mean subtraction was shown, with the improved performance being linked to unwanted subject variabilities being removed. An interesting point was also raised about the validity of using a standard HMM for speech recognition in the visual modality, as the quasi stationary assumption made for the acoustic modality does *not* seem to hold as well in the visual modality. Visual speaker recognition achieved excellent results using the SLDA mouth feature. This can be attributed to the more static nature of the speaker recognition task, which is easily accommodated by the LDA feature extraction procedure and standard HMM topology.

References

1. F. Lavagetto, "Converting speech into lip movements: A multimedia telephone for hard hearing people," *IEEE Transactions on Rehabilitation Engineering*, vol. 3, pp. 90–102, March 1995.
2. G. Potamianos, H. P. Graf, and E. Cosatto, "An image transform approach for HMM based automatic lipreading," in *International Conference on Image Processing*, vol. 3, pp. 173–177, 1998.
3. S. Young, D. Kershaw, J. Odell, D. Ollason, V. Valtchev, and P. Woodland, *The HTK Book (for HTK version 2.2)*. Entropic Ltd., 1999.
4. C. Bregler and Y. Konig, "Eigenlips for robust speech recognition," in *International Conference on Accoustics, Speech and Signal Processing*, (Adelaide, Australia), 1994.
5. G. Potamianos and H. P. Graf, "Linear discriminant analysis for speechreading," in *IEEE Second Workshop on Multimedia Signal Processing*, pp. 221–226, 1998.
6. G. Potamianos, J. Luettin, and C. Neti, "Hierarchical discriminant features for audio-visual LVCSR," in *International Conference on Acoustics, Speech and Signal Processing*, vol. 1, pp. 165–168, 2001.
7. M. S. Gray, J. R. Movellan, and T. J. Sejnowski, "A comparison of local versus global image decompositions for visual speechreading," in *4th Joint Symposium on Neural Computation*, pp. 92–98, 1997.
8. I. Matthews, T. Cootes, S. Cox, R. Harvey, and J. A. Bangham, "Lipreading using shape, shading and scale," in *Auditory-Visual Speech Processing*, (Sydney, Australia), pp. 73–78, 1998.
9. S. Pigeon, "The M2VTS database," (Laboratoire de Telecommunications et Teledection, Place du Levant, 2-B-1348 Louvain-La-Neuve, Belgium), 1996.
10. S. Cox, I. Matthews, and J. A. Bangham, "Combining noise compensation with visual Information in speech recognition," in *Auditory-Visual Speech Processing*, (Rhodes), 1997.
11. L. Deng, "A generalized hidden Markov model with state-conditioned trend functions of time for the speech signal," *Signal Processing*, vol. 27, pp. 65–78, 1992.