IMPROVED SPEAKER INDEPENDENT LIP READING USING SPEAKER ADAPTIVE TRAINING AND DEEP NEURAL NETWORKS

Ibrahim Almajai, Stephen Cox, Richard Harvey, Yuxuan Lan

School of Computing Sciences, University of East Anglia, Norwich, NR7 7TJ.

i.almajai@gmail.com, s.j.cox, r.w.harvey, y.lan@uea.ac.uk

ABSTRACT

Recent improvements in tracking and feature extraction mean that speaker-dependent lip-reading of continuous speech using a medium size vocabulary (around 1000 words) is realistic. However, the recognition of previously unseen talkers has been found to be a very challenging task, because of the large variation in lip-shapes across speakers and the lack of large, tracked databases of visual features, which are very expensive to produce. By adapting a technique that is established in speech recognition but has not previously been used in lip-reading, we show that error-rates for speaker-independent lip-reading can be very significantly reduced. Furthermore, we show that error-rates can be even further reduced by the additional use of Deep Neural Networks (DNNs). We also find that there is no need to map phonemes to visemes for context-dependent visual speech transcription.

Index Terms— Lipreading, Deep Neural Networks, Speaker Adaptive Training

1. INTRODUCTION

Automatic lip-reading is known to be a difficult problem. So far, the technology of automatic lip-reading has been largely confined to constrained tasks such as: small vocabulary recognition [1] where the number of words is constrained; talker-dependent recognition where the number of talkers is constrained, or it has been relegated to a means of boosting the performance of conventional audio speech recognition (audio-visual recognition[2]). Furthermore, it is also notoriously difficult to measure human performance on lip-reading tasks, but the few studies that exist indicate that even hearing-impaired people achieve rather low word accuracy rates when lip-reading speakers they have never seen before [3]. Anecdotally, hearing-impaired people tell us that different people are easier or harder to lip-read, depending on matters such as their physiognomy, how much effort they put into articulation during speech, and even the presence of facial hair.

Speaker-independent recognition has not been studied very much. In [4], the authors present results for a ten isolated word speaker-independent system, and we examined speaker-independent recognition on a 1090-word database of continuous English speech derived from the Resource Management (RM) database [5]. This work has been complemented by developments in tracking and feature extraction: [6] demonstrated that tracking and feature extraction are possible even on outdoor scenes with video taken by hand-held domestic interlaced cameras. The situation is akin to earlier work on acoustic recognition—many of the problems that were thought to be difficult are being solved, leaving the most difficult one, which is that of speaker-independent recognition.

Recently, Deep Neural Networks (DNN) and some other deep learning architectures have proved to be successful in Automatic Speech Recognition (ASR) and other areas of machine learning [7]. A lot of research has already been published in which deep learning techniques are applied to ASR. However, much less work has been done on applying those tecnhiques to automatic lip-reading. Ngiam et al. [8] applied unsupervised deep learning to learn cross modality features of audio and video speech data. The first stage of training is Restricted Boltzman Machines (RBMs) to unsupervisedly learn a better representaion of audio and visual features. The learned features are then passed to a deep autoencoder where training is supervised. They reported a classification improvement in some tasks when only visual features are available at supervised training and testing but both madalities are present at the feature learning stage.

Maximum Likelihood Linear Transform (MLLT) is a standard technique in ASR [9] and has also been applied to Audio-Visual Speech Recognition (AVSR) [10]. In MLLT, the idea is to find a linear transform of the input features in which the assumption of a diagonal covariance matrix is the most valid (in the sense of loss of likelihood compared with using full covariance matrices). When this condition is met, modelling is closer to using full covariance matrices and it can be shown that inter-class discrimination is improved.

Previous work has shown that the features obtained from the lips are highly speaker-dependent [11]. In this paper we show that the application of Speaker Adaptive Training (SAT), which is also a standard technique in ASR, appears to have considerable promise in speaker-independent lip-reading. SAT is a technique for normalising the effects of variation in the acoustic features of different speakers when

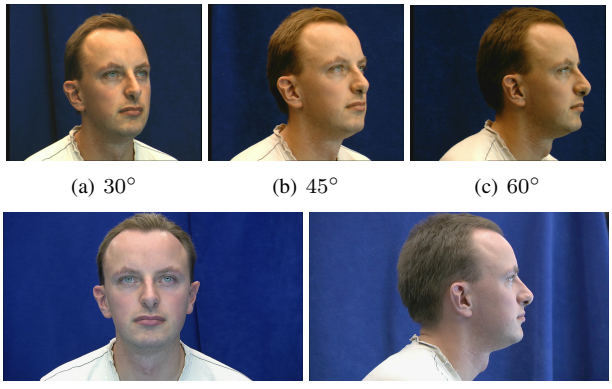
training a set of acoustic models for recognition. It basically avoids modelling the inter-speaker variability and only models the intra-speaker variability. Individual speaker characteristics are modelled by linear transformations of the mean parameters of the acoustic models. The algorithm functions by alternately optimising the model means and the transformation parameters for a particular speaker.

We report the best known results for speaker-independent lip-reading by using a combination of MLLT followed by SAT. We also report the performance of a "hybrid" Context-Dependent Deep Neural Networks (CD-DNNs) where Context-Dependent Gaussian mixture model (CD-GMM) likelihoods in HMMs are replaced by posterior probablities of DNNs after being converted into quasi-likelihoods [12].

The result is useful because it first challenges the conventional wisdom that speaker-independent recognition is not viable. Second, it shows DNNs to be promising for speaker-independent lip-reading despite the limited amount of training data.

2. DATASET AND FEATURES

For data we use an audiovisual corpus of twelve speakers, seven male and five female, each reciting 200 sentences selected from the Resource Management Corpus[5]. Figure 1 shows an example of the data which was recorded on five gen-locked cameras from different angles. Here we use only



(a) 30° (b) 45° (c) 60°

(d) 0° (e) 90°

Fig. 1. Different views of LiLIR dataset[13]

the front view, which was recorded using a tri-chip Thomson Viper FilmStream high-definition camera at a resolution of 1920×1080. The database has a vocabulary size of 1090 words and consists of a number of stylized sentences such as "Give me Constellation's displacement in long tonnes ". Previous tests on these data with professional human lipreaders [3] revealed viseme error rates of 39.7% to 85.4% and word-error rates of 0% to 69% (compared to a viseme accuracy of 46% and a word accuracy of 14% for the automatic system).

Each video has been tracked using linear-predictor based tracker (described in [14]). To generate AAM features, an Active Appearance Model (as in [3]) was trained using an one-held-out methodology (that is, the model used to describe speaker n was trained using all speakers except speaker n). In previous work we have examined several choices of features that appeared to work best with a Hidden-Markov Model based classifier implemented using the Hidden Markov Model Toolkit (HTK)[15]. Among the best features were the combined Shape and Appearance model (denoted CSAM in [13]). In these features, the shape vector, $\mathbf{s}$ and the appearance vector $\mathbf{a}$ are further combined using PCA to produce a combined feature vector. Here we retain 97% of the variation and the combined feature size is typically 21- or 22-dimensional.

3. EXPERIMENTS

Kaldi speech recognition toolkit [16] was used to train our visual speech models (phonemes and visemes units) and decode the test data. Visemes are visually distinguishable speech units which have a one-to-many mapping to phonemes. Fisher phoneme-to-viseme mapping [17][18] is used and shown in Table 1.

Table 1. Fisher mapping of 45 phonemes to 14 visemes including silence

Viseme	Phonemes
V1	/b/ /p/ /m/
V2	/f/ /v/
V3	/t/ /d/ /s/ /z/ /th/ /dh/
V4	/w/ /r/
V5	/k/ /g/ /n/ /l/ /ng/ /hh/ /h/ /y/
V6	/ch/ /jh/ /sh/ /zh/
V7	/eh/ /ey/ /ae/ /aw/ /er/ /ea/
V8	/uh/ /uw/
V9	/iy/ /ih/ /ia/
V10	/ah/ /ax/ /ay/
V11	/ao/ /oy/ /ow/ /ua/
V12	/aa/
V13	/oh/
V14	/sil/

The HMM/GMM systems we built are: (i) monophone and monoviseme systems with Δ and $\Delta\Delta$ features (Mono), (ii) triphone and triviseme systems with LDA ((Tri:LDA)) (iii) triphone and triviseme systems with LDA+MLLT ((Tri:LDA+MLLT)), (iv) triphone and triviseme systems with LDA+MLLT+SAT (Tri:LDA+MLLT+SAT). The visual features are also augmented by their velocity and acceleration temporal derivative

The feature processing pipeline up to the DNN stage is summarised in Figure 2. Firstly the visual features are considered in a block of 7 frames (termed "splicing" in [12]).

They are then decorrelated and forced to a dimensionality of 40 using Linear Discriminant Analysis (LDA) and further decorrelated using maximum likelihood linear transform (MLLT)[9]. Speaker Adaptive Training (SAT)[19] is then applied using feature-space maximum likelihood linear regression (fMLLR) of 40×41. The 40-dimensional speaker adapted features are then spliced across a window of 9 frames and applying LDA to decorrelate the concatenated features and reduce dimensionality to 250 [12].

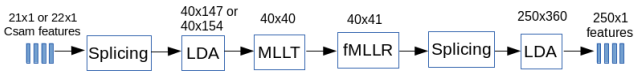


Fig. 2. Schematic diagram of the feature processing

Speaker Adaptive Training (SAT)[19] is then applied using feature-space maximum likelihood linear regression (fMLLR) of 40×41. The 40-dimensional speaker adapted features are then spliced across a window of 9 frames and applying LDA to decorrelate the concatenated features and reduce dimensionality to 250 [12].

Note that in the first-stage of the configuration, the learning is conventional, using an HMM with Gaussian Mixture Models for modelling the output. In the case of mono-models, a total number of Gaussians is chosen to be 600. For training the other three types of context-dependent models, a maximum number of leaves for the decision tree is 700 and total number of Gaussians is 3000.

The DNN had four hidden layers, where the output layer is a soft-max layer and corresponds to a maximum of 700 states of context-dependent HMM states. The hidden layers contain sigmoid nonlinearities which have a range of $(0, 1)$ and the connection weights were randomly initialised with a normal distribution multiplied by 0.1 [12, 16]. DNN is trained to classify frames into context-dependent states by mini-batch Stochastic Gradient Descent. The learning rate was intially set to 0.02 and kept fixed during the rest of 15 epochs as long as the increment in cross-validation frame accuracy in a single epoch was higher than 0.5%. If not, the learning rate was halved; this was repeated until it was less than 0.004. The weights are updated using mini-batches of size 64 frames. The total number of DNN parameters is 1 million. The alignment of context-dependent states to frames derives from the GMM stage (LDA+MLLT+SAT) and is left fixed during training. The DNN experiments use conventional CPUs rather GPUs [12, 16].

4. RESULTS

Figure 3 shows the word accuracy results for each of the twelve speakers tested on our system using viseme units, with the mean performances shown as the final column. The "Mono" results were made using a single model of each viseme followed by a word-bigram model. Moving to tri-visemes (the visual equivalent of triphones in audio recognition) increases the number of potential classes but there is a significant increase word accuracy.

The four tri-viseme configurations are LDA (which is the first two boxes of Figure 2), LDA plus MLLT (the first htree boxes of Figure 2) and LDA + MLLT + SAT (all the boxes in Figure 2). Also shown are the results using DNNs.

Figure 3 shows that, with very few exceptions, performance increases with each stage for each speaker. Sometimes the gain is small (typically when adding MLLT to the LDA features) but some stages show larger gains.

The mean results across all speakers are summarised in Figure 3. Word recognition accuracy is always higher when phonemes are used as the modelling units rather than visemes. This confirms what has been recently established on a speaker dependent task [18]. This is counter-intuitive, since many of the features that distinguish phonemes can't be seen (e.g. voicing, or place of articulation when it is far back in the mouth). However, the viseme to phoneme mapping introduces ambiguity: because it is a many-to-one mapping, some words have the same visemic transcription (*homophenous* words) [18].

The largest performance increase appears to come from the addition of SAT. This is satisfying, because previous work [11], showed that the visual features that we use that represent a certain sound are highly speaker-dependent, and hence this feature normalisation by speaker is highly beneficial. It is also worth noting that the amount of training data is rather small for the DNN stage, so we think the DNNs have more potential performance.

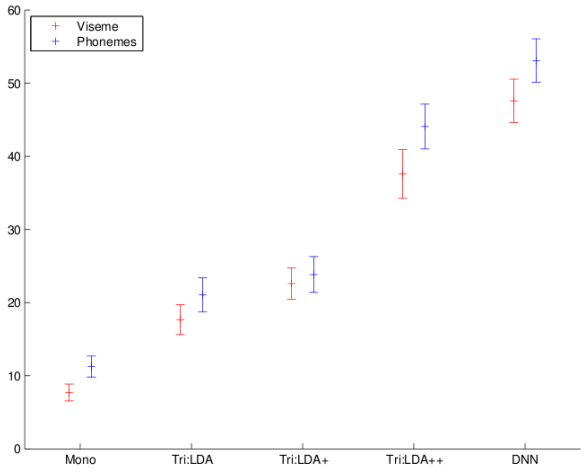


Fig. 4. Phoneme- and viseme-based recognisers compared showing mean word accuracy ± 1 standard error

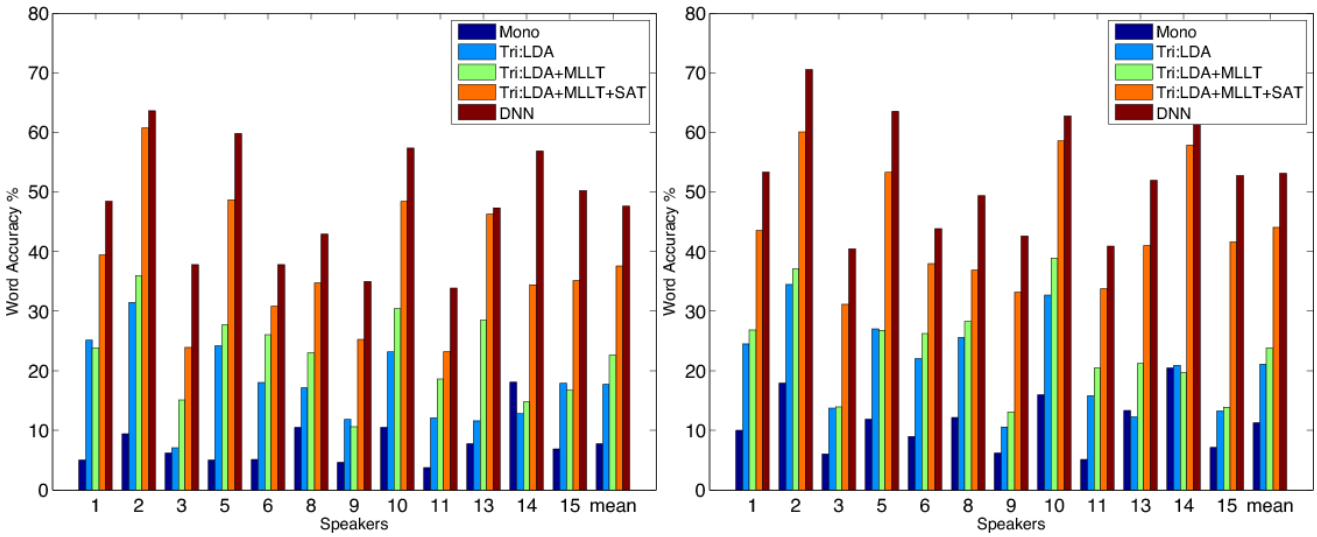


Fig. 3. Word accuracy for various talkers using Type IV features. Left: recognition using visemes as units. Right: phonemes.

5. CONCLUSIONS

Speaker-independent recognition has been seen as an unachievable goal of lip-reading for sometime. Even skilled human lip-readers find that their performance is high talker-dependent [3, 20]. In this paper we added the stage of SAT, which is an essential technique in the state-of-art acoustic speech recognition system, to classify a medium-sized vocabulary visual speech dataset of continuous speech.

Our results indicate that speaker-independent lip-reading is indeed viable, even with relatively small amounts of training data and there is considerable potential for further improvement. More training data will certainly improve results, but it labelled video data collection is still an expensive exercise—one possible improvement would be to collect video data from opportunistic sources such as TV and video. Another intriguing finding is the difference in performance between phoneme and viseme units. The majority of lip-reading systems are constructed around one of number of standard viseme sets, but it appears that ignoring these *ad hoc* mappings leads to better results. At the moment there are number of possible explanations for this effect but this is also an area for future work since it raises the possiblity of alternative viseme sets or a challenge to the conventional wisdom about what can and cannot be seen on the lips.

6. REFERENCES

[1] 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, February 2002.

[2] G. Potamianos, C. Neti, J. Luettin, and I. Matthews, "Audio-visual automatic speech recognition: An overview," in *Issues in Visual and Audio-visual Speech Processing*. 2004, MIT Press.

[3] Yuxuan Lan, Richard Harvey, and Barry-John Theobald, "Insights into machine lip reading," in *Acoustics, Speech and Signal Processing (ICASSP), 2012 IEEE International Conference on*, March 2012, pp. 4825 – 4828.

[4] Ziheng Zhou, Xiaopeng Hong, Guoying Zhao, and M. Pietikainen, "A compact representation of visual speech data using latent variables," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 36, no. 1, pp. 1–1, Jan 2014.

[5] W.M. Fisher, G.R. Doddington, and K.M. Goudie-Marshall, "The DARPA speech recognition research database: specifications and status.," in *In Proceedings of the DARPA Speech Recognition Workshop.*, 1986.

[6] Richard Bowden, Stephen Cox, Richard Harvey, Yuxuan Lan, Eng-Jon Ong, Gari Owen, and Barry-John Theobald, "Recent developments in automated lip-reading," 2013, vol. 8901, pp. 89010J–89010J–13.

[7] Geoffrey Hinton, Li Deng, Dong Yu, George E Dahl, Abdel-rahman Mohamed, Navdeep Jaitly, Andrew Senior, Vincent Vanhoucke, Patrick Nguyen, Tara N Sainath, et al., "Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups," *Signal Processing Magazine, IEEE*, vol. 29, no. 6, pp. 82–97, 2012.

[8] Jiquan Ngiam, Aditya Khosla, Mingyu Kim, Juhan Nam, Honglak Lee, and Andrew Y Ng, "Multimodal

deep learning," in *Proceedings of the 28th international conference on machine learning*, 2011, pp. 689–696.

[9] Ramesh A Gopinath, "Maximum likelihood modeling with gaussian distributions for classification," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, 1998, vol. 2, pp. 661–664.

[10] Gerasimos Potamianos, Chalapathy Neti, Guillaume Gravier, Ashutosh Garg, and Andrew W. Senior, "Recent advances in the automatic recognition of audio-visual speech," in *Proceedings of the IEEE*, Sept 2003, vol. 91, pp. 1306–1326.

[11] Stephen Cox, Richard Harvey, Yuxuan Lan, Jacob Newman, and Barry-John Theobald, "The challenge of multispeaker lip-reading," in *International Conference on Auditory-Visual Speech Processing*. Citeseer, 2008, pp. 179–184.

[12] Shakti P Rath, Daniel Povey, K Vesely, and J Cernocky, "Improved feature processing for Deep Neural Networks," in *Proc. of Interspeech*, August 2013.

[13] Yuxuan Lan, Barry-John Theobald, and Richard Harvey, "View independent computer lip-reading," in *IEEE Conference on Multimedia and Expo (ICME 2012)*. July 2012, IEEE.

[14] E. Ong, Y. Lan, B. Theobald, R. Harvey, and R. Bowden, "Robust facial feature tracking using selected multi-resolution linear predictors," in *In Proceedings of the International Conference Computer Vision (ICCV)*, 2009.

[15] Steve Young, Gunnar Evenmann, Dan Kershaw, Gareth Moore, Julian Odell, Dave Ollason, Valtcho Valtchev, and Phil Woodland, *The HTK Book (version 3.2.1)*, 2002.

[16] Daniel Povey, Arnab Ghoshal, Gilles Boulianne, Lukas Burget, Ondrej Glembek, Nagendra Goel, Mirko Hannemann, Petr Motlicek, Yanmin Qian, Petr Schwarz, et al., "The Kaldi speech recognition toolkit," in *Proc. ASRU*, 2011, pp. 1–4.

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

[18] Dominic Howell, *Confusion modelling for lip-reading*, Ph.D. thesis, University of East Anglia, 2015.

[19] Tasos Anastasakos, John McDonough, and John Makhoul, "Speaker adaptive training: A maximum likelihood approach to speaker normalization," in *Acoustics, Speech, and Signal Processing, 1997. ICASSP-97., 1997 IEEE International Conference on*. IEEE, 1997, vol. 2, pp. 1043–1046.

[20] DeborahA. Yakel, LawrenceD. Rosenblum, and MichelleA. Fortier, "Effects of talker variability on speechreading," *Perception & Psychophysics*, vol. 62, no. 7, pp. 1405–1412, 2000.