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 speakers 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 (DNN). We also find that there is no need to map phonemes to visemes for context-dependent visual speech transcription.

Index Terms— Automatic lip-reading, 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] [2] [3] where the number of words is constrained; speaker-dependent recognition where the number of speakers is constrained, or it has been relegated to a means of boosting the performance of conventional audio speech recognition (audio-visual speech recognition [4]). Furthermore, the few studies that exist on the difficult task to measure human lip-reading performance indicate that even hearing-impaired people achieve rather low word accuracy rates when lip-reading speakers they have never seen before [5].

Speaker-independent lip-reading has not been studied very much. [6] presents results for a ten isolated word speaker-independent system. In [3], the authors investigate speaker-dependent, multi-speaker and speaker-independent lip-reading using two isolated letters datasets (AVletters and AVletters 2). They show significant performance drop in speaker-independent recognition tasks compared to the other two configurations and find that the use of the Maximum Likelihood Linear Regression (MLLR) adaptation technique is not sufficient to compensate for the drop in performance. In this work, we examine speaker-independent recognition on around 1000 words database of continuous English speech derived from the Resource Management (RM) corpus [7]. The work has been complemented by developments in tracking and feature extraction: [8] demonstrated that tracking and feature extraction are possible even on outdoor scenes with video taken by hand-held domestic interlaced cameras.

Recently, Deep Neural Networks (DNN) with different deep learning architectures have proved to be successful in Automatic Speech Recognition (ASR) and other areas of machine learning [9]. 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 techniques to automatic lip-reading. Some research has been published on Audio-Visual Speech Recognition (AVSR) using deep learning [10] [11] [12] [13] [14] [15]. [10] 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 representation of audio and visual features. The learned features are then passed to a deep autoencoder where training is supervised. They reported a classification improvement on AVletters and CUAVE when only visual features are available at supervised training and testing but both modalities are present at the feature learning stage. In [15], the use of a context-dependent DNN system on a single speaker dataset, (RM)-3000, gave a word accuracy of 85% with a 33% improvement on the baseline HMM.

The Maximum Likelihood Linear Transform (MLLT) is a standard technique in ASR [16] and has also been applied to AVSR [17] [15]. 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 [3]. 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-DNN) where Context-Dependent Gaussian mixture model (CD-GMM) likelihoods in HMM are replaced by posterior probabilities of DNN after being converted into quasi-likelihoods [18].

The result is useful because it first challenges the conventional wisdom that speaker-independent recognition is extremely difficult. Second, it shows DNN to be promising for speaker-independent lip-reading despite the limited amount of training data and without the inclusion of a pre-training stage (feature learning).

2. DATASET AND FEATURES

For data, we use an audiovisual corpus of twelve speakers [19], seven male and five female, each reciting 200 sentences selected from the RM corpus [7]. The vocabulary size is approximately 1000 words. Figure 1 shows an example of the data which was recorded on five gen-locked cameras from different angles. Here we use only the front view, which was

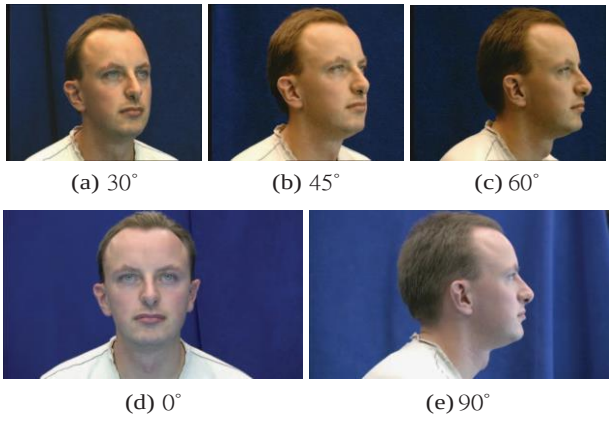

Fig. 1. Different views of the dataset [20]

recorded using a tri-chip Thomson Viper FilmStream high-definition camera at a resolution of 1920×1080. The database has a vocabulary size of around a 1000 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 lip-readers [5] 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 [21]). To generate AAM features, an Active Appearance Model (as in [5]) 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) [22]. Among the best features were the combined Shape and Appearance model (denoted CSAM in [20]). 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 [23] was used to train our visual speech models (phonemes and visemes units) and decode the test data using a strategy of 12-fold cross-validation: for each fold, a different speaker is held-out for testing and the classifier's models are trained on the data of the remaining speakers. Visemes are visually distinguishable speech units which have a one-to-many mapping to phonemes. Fisher phoneme-to-viseme mapping [24] [25] 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 ΔΔ features, (ii) triphone and triviseme systems with LDA, (iii) triphone and triviseme systems with LDA+MLLT, (iv) triphone and triviseme systems with LDA+MLLT+SAT. Kaldi's automatic method of building decision trees without the need to provide a set of questions [23] is quite convenient to build context-dependent lip-reading systems. To compose the Kaldi decoding graph, a word-pair bigram language model was built and the lexicon was derived from the RM distribution [7].

The feature processing pipeline up to the DNN stage is summarised in Figure 2. Firstly the visual features are mean-normalised on a per-speaker basis before considered in a block of 7 frames. 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) [16]. Speaker Adaptive Training (SAT) [26] 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 [18]. The fMLLR is also applied to the features of the test speaker. For DNN only, 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 [18].

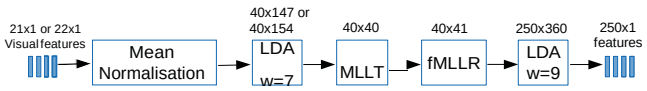

Fig. 2. Schematic diagram of the feature processing where *w* represents the window width of LDA.

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 number of Guassians per HMM state is decided automatically based data count [23]. The actual number of leaves after clustering is always lower than the given maximum, for instance, it ranges from 512 to 544 in the 12 experiments of the triphones systems. The decoding beam length is 30 and lattice beam width is 10.

The DNN system is trained using the alignment of context-dependent states and the decision tree derived from the GMM stage (LDA+MLLT+SAT). It had four hidden layers of tahn units and the output layer is a soft-max layer of size 2000. This size was deliberately chosen to be larger than the number of leaves in the decision tree to allow the creation of multiple "virtual" targets for each leaf [23]. Training the DNN utilised the mini-batch stochastic gradient descent technique. The weights were updated using mini-batches of size 64 frames. The total number of DNN parameters was chosen to be 1 million which made the number of the tanh units in each of the four hidden layers to be 491. The learning rate was initially 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 decoding beam length was 30 and lattice beam width was 18. The DNN experiments use conventional CPUs rather than GPUs [18, 23].

4. RESULTS

Figure 4 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. Moving to trivisemes increases the number of potential classes but there is a significant increase word accuracy. The four triviseme configurations are LDA (which is the first two boxes of Figure 2), LDA plus MLLT (the first three boxes of Figure 2) and LDA + MLLT + SAT (all the boxes in Figure 2). Also shown are the results using DNN.

Figure 3 shows that, with very few exceptions, performance increases with each stage for every 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 4. 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 [25]. 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) [25].

The largest performance increase appears to come from the addition of SAT. This is satisfying, because previous work [3], showed that the visual features that we use that represent a certain sound are highly speaker-dependent, and hence this feature adaptation 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.

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 speaker-dependent [5, 27]. In this paper, we incorporated SAT and fMLLR, which are essential techniques in the state-of-art

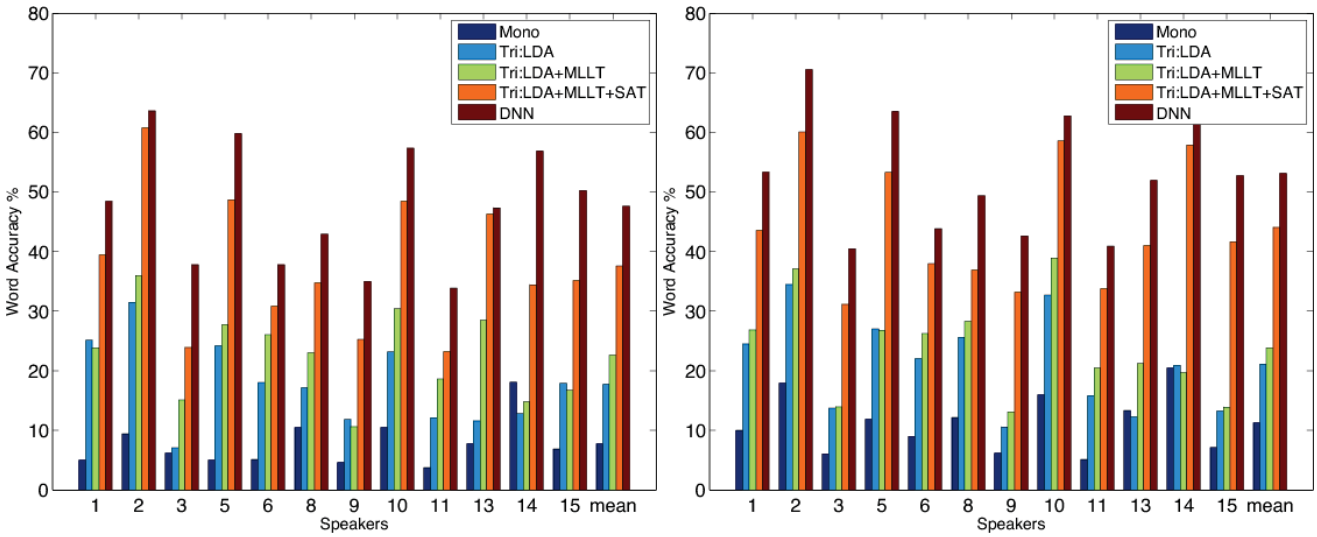

Fig. 3. Word accuracy for various speakers using Type IV features. Left: recognition using visemes as units. Right: phonemes.

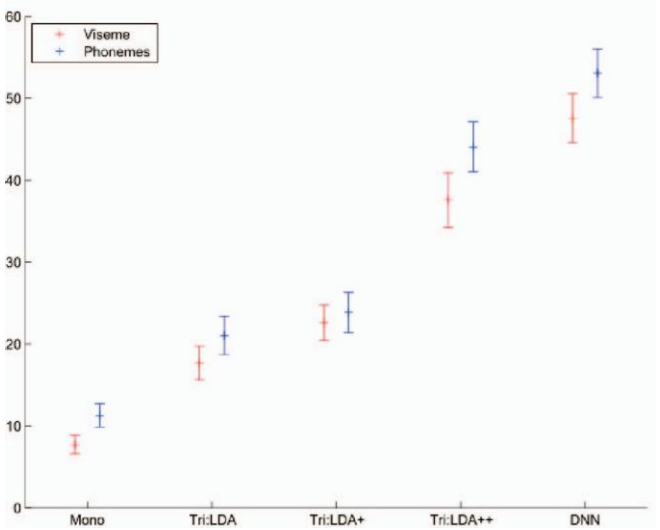

Fig. 4. Phoneme- and viseme-based recognisers compared showing mean word accuracy ±1 standard error

ASR, in the processing pipeline 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. Although the majority of lip-reading systems are constructed around one of number of standard viseme sets, it appears that ignoring these *ad hoc* mappings leads to better results even in speaker-independent tasks.

Given the increased evidence that the use of DNN is beneficial in lip-reading, future work should investigate the use of more of the different DNN architectures and training strategies available and applied to ASR.

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] Shi-Lin Wang, Wai H Lau, Alan Wee-Chung Liew, and SH Leung, "Automatic lipreading with limited training data," in *Pattern Recognition, 2006. ICPR 2006. 18th International Conference on*. IEEE, 2006, vol. 3, pp. 881–884.

[3] 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.

[4] 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.

[5] 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.

[6] 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.

[7] P Price, WM Fisher, J Bernstein, and DS Pallett, "Resource management rm1 2.0," *Linguistic Data Consortium, Philadelphia*, 1993.

[8] 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.

[9] 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.

[10] 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.

[11] Jing Huang and Brian Kingsbury, "Audio-visual deep learning for noise robust speech recognition," in *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*. IEEE, 2013, pp. 7596–7599.

[12] Seungwhan Moon, Suyoun Kim, and Haohan Wang, "Multimodal transfer deep learning for audio visual recognition," *arXiv preprint arXiv:1412.3121*, 2014.

[13] Youssef Mroueh, Etienne Marcheret, and Vaibhava Goel, "Deep multimodal learning for audio-visual speech recognition," *arXiv preprint arXiv:1501.05396*, 2015.

[14] Kuniaki Noda, Yuki Yamaguchi, Kazuhiro Nakadai, Hiroshi G Okuno, and Tetsuya Ogata, "Audio-visual speech recognition using deep learning," *Applied Intelligence*, vol. 42, no. 4, pp. 722–737, 2015.

[15] Kwanchiva Thangthai, Richard Harvey, Stephen Cox, and Barry-John Theobald, "Improving lip-reading performance for robust audiovisual speech recognition using dnns," 2015.

[16] 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.

[17] 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.

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

[19] Yuxuan Lan, Barry-John Theobald, Richard Harvey, Eng-Jon Ong, and Richard Bowden, "Improving visual features for lip-reading.," in *AVSP*, 2010, pp. 7–3.

[20] 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.

[21] 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.

[22] 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.

[23] 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.

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

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

[26] 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.

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