Extraction of Features for Lip-reading Using Autoencoders

Karel Paleček

The Institute of Information Technology and Electronics,
Technical University of Liberec, Studentská 2/1402, 46117 Liberec, Czech Republic
karel.palecek@tul.cz

Abstract. We study the incorporation of facial depth data in the task of isolated word visual speech recognition. We propose novel features based on unsupervised training of a single layer autoencoder. The features are extracted from both video and depth channels obtained by Microsoft Kinect device. We perform all experiments on our database of 54 speakers, each uttering 50 words. We compare our autoencoder features to traditional methods such as DCT or PCA. The features are further processed by simplified variant of hierarchical linear discriminant analysis in order to capture the speech dynamics. The classification is performed using a multi-stream Hidden Markov Model for various combinations of audio, video, and depth channels. We also evaluate visual features in the join audio-video isolated word recognition in noisy environments. English

Keywords: Autoencoder, Hidden Markov Model, Kinect, Lip-reading.

1 Introduction

Automatic visual speech recognition, or lip-reading, has been an active research area for over two decades now. Many studies conducted over the time demonstrated an improved accuracy when incorporating visual information over audio-only speech recognition) [1], [2], [3], especially in noisy environments.

Existing lip-reading methods may be broadly classified into two main groups: methods solely exploiting appearance-based features and methods modeling shape as well. For the appearance-based methods, visual features are extracted from a region of interest (ROI), usually a rectangular area centered around the speaker's mouth. Some of the most commonly used features are Discrete Cosine Transformation (DCT) and Principal Component Analysis (PCA) [4]. For the second class of algorithms, shape can be represented using e.g. a set of several facial landmarks and then modeled by multivariate distributions. Examples of such methods include e.g. Active Appearance Model (AAM) [4], [5]. While the combined shape and appearance methods usually perform better, they require stable and reliable landmark detection, which is a non-trivial task. In this work, we extract the visual features from the ROI, not modeling the shape of the speaker's lips.

Most of the research has been focused on extracting visual cues from the frontal image of the speaker's face, therefore not modeling 3D properties of the

A. Ronzhin et al. (Eds.): SPECOM 2014, LNAI 8773, pp. 209–216, 2014.

ROI. There have been studies where the authors recorded speakers by multiple cameras and then performed lip-reading using 3D information reconstructed by stereo-vision algorithms [6], [7]. However, due to sensitivity to lighting conditions, hardware requirements, and computational complexity, these methods remain rather scarcely used in the context of visual speech recognition. In the recent years, few affordable devices such as Asus Xtion, Creative Senz3D or Microsoft Kinect have become popular for 3D reconstruction. These devices are able to reconstruct depth information using structured light and depth from focus techniques. One of the pioneering efforts in lip-reading with incorporating facial depth data from Kinect was [8], where the authors applied 2D DCT to both video and depth streams, and combined the modalities via multi-stream hidden Markov model. In [9], patch trajectories were extracted from video and depth, and used in a random forest manifold alignment algorithm for lipreading.

Recently, a class of methods known as deep learning has gained an increased attention in the computer vision and speech recognition communities. Deep learning algorithms are most commonly used as an unsupervised pre-training procedure that automatically extracts useful information from the data for deep neural network classification. This is done in a greedy layer-wise manner by fitting e.g. Restricted Boltzmann Machine (RBM) or an autoencoder for each layer. For an overview, see [10]. In [11] and [12], authors used deep neural networks, pre-trained on several concatenated PCA-reduced frames using RBM, for visual speech parametrization and achieved better results than using hand-engineered features.

In our work, we propose a single layer only autoencoder as a feature extraction method for visual speech recognition, and use it to extract features from both video and depth streams. In contrast to [11] and [12], we apply the autoencoder directly on the image and not on concatenated feature vectors. Instead, similarly to [8], we incorporate speech dynamics on higher-level features and classify using a multi-stream Hidden Markov Model. In the experiments on our database recorded by Kinect, we demonstrate the improved performance of the autoencoder features over DCT and PCA in the task of isolated word recognition with incorporated facial depth data.

2 Feature Extraction

An autoencoder [10], also called an autoassociator, is a type of neural network that learns a distributed representation of the input. In consists of two parts: an encoder that converts the input into activations of its hidden units, and decoder that reconstructs the input from the encoder's internal representation. The input vector $x \in \mathbb{R}^n$ is encoded by m hidden units as

$$h(x) = f\left(Wx + b\right) \tag{1}$$

where W is $m \times n$ matrix of weights of each unit, b is a $m \times 1$ bias vector, and $f(\cdot)$ is an element-wise activation function. If $m < n$ and f is linear, it can be shown that the learned representation $h(x) \in \mathbb{R}^m$ lies in the subspace of eigenspace of

the input data. In order to find more interesting features that are better suited for classification, we use the sigmoid activation, i.e. $f(z) = \sigma(z) = \frac{1}{1+\exp(-z)}$. The input is then reconstructed by the decoder as

$$y = f\left(W'h(x) + c\right) \tag{2}$$

where W' is $n \times m$ matrix of decoder's connection weights and c is a $n \times 1$ bias vector. In our work, we consider autoencoder with tied weights, i.e. where $W' = W^\top$. The aim of the autoencoder is to learn W, b, and c such that a reconstruction error $L(x, y)$ is minimized. Since we deal with real-valued data, we define the reconstruction error as

$$L(x, y) = \|y - x\|^2 + \alpha \sum_{ij} w_{ij}^2 \tag{3}$$

The regularization term in (3) prevents overfitting by keeping the weights w_{ij} of the matrix W small. From the probabilistic perspective this corresponds to imposing a Gaussian prior on the weights w_{ij}.

In order to limit the number of images for which each neuron is active (i.e. its output value is close to 1), we apply additional L_1 regularization penalty on the input to the sigmoid function. The complete objective function of our autoencoder therefore takes the form

$$J\left(W, b, c\right) = \frac{1}{|X|} \sum_{x \in X} \|y - x\|^2 + \alpha \sum_{ij} w_{ij}^2 + \beta \sum_{x \in X} \sum_{j=1}^{m} \left|w_i^\top x\right| \tag{4}$$

where $w_i^\top$ is the i-th row of the matrix W. We find the optimal W, b, and c by minimizing (4) with respect to using the Limited memory Broyden-Fletcher-Goldfarb-Shanno (L-BFGS) algorithm. The weights of each unit are initialized to small uniformly distributed random values inversely proportional to the total number of its connections. The bias vectors b and c are initialized to zeros. After the optimal W, b, and c has been found, the hidden representation (1) is used as a visual speech parametrization vector.

Since optimization of the objective function (4) is a computationally expensive task, tuning the hyper-parameters α and β using exhaustive grid search techniques is not feasible, because the number of experiments would be too large. Therefore, in order to find the optimal values for α and β, we employ Bayesian optimization strategy with the expected improvement acquisition function [18]. Bayesian optimization is a general method for minimization of an unknown function. It utilizes Monte-Carlo techniques to select each evaluation point in the parameter space. In our case, the objective function is defined as the word error rate (WER) that is achieved by classifying the features learned by the autoencoder. The classification is performed using a whole-word Hidden Markov Model (HMM) on the full cross validated database. Examples of features learned by our autoencoder (AE) are shown in Fig. 1. Note that some AE features fail to converge.

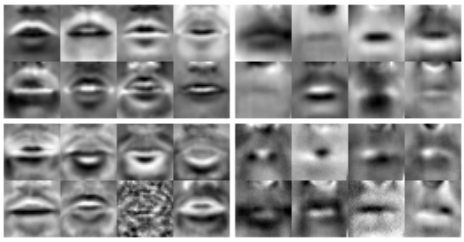

Fig. 1. Examples of learned features. Top row: PCA features, bottom row: AE. Left: video, right: depth channel.

3 Visual Front-End

Region of interest extraction and data preprocessing consists of several stages in our work. In the first stage, position of the speakers face is approximately estimated using Viola-Jones detector (VJ) [13], a well known method based on computationally inexpensive Haar-like features that are combined into a strong classifier using boosting technique.

In the second stage, location and shape of speakers lips, chin and mouth are refined using Explicit Shape Regression algorithm (ESR) [14]. Similarly to traditional face alignment methods such as Active Appearance Model (AAM), ESR models shape of an object by set of N landmarks, but instead of modeling complex distributions of the shape variance, it predicts optimal joint landmark configuration discriminatively based on the current estimate. However, since there is no objective function to be minimized during the face alignment stage, the predicted facial shape is slightly different in each frame, causing random noise in the position of the landmarks. In order to extract the region of interest (ROI) in a more stable way, we therefore average the fitting results over three neighboring frames in time.

In our work, the region of interest (ROI) is defined as square region covering the mouth and its closest surroundings. The scale invariance is achieved by defining the ROI on the unit-normalized mean facial shape obtained by aligning and then averaging all shapes in the training database. For each frame, the mean facial shape is aligned to the detected shape by Euclidean transformation such that the mean square error is minimized. The size of the ROI is fixed to 32×32 pixels because of efficiency reasons of the auto-encoder fitting procedure.

4 Data Preparation

For experimenting with visual speech features incorporating depth information we recorded an audio-visual database containing both isolated word and continuous speech utterances. Our database contains 54 speakers (23 female and 31 male), each uttering 50 isolated words in Czech language. The database also contains 583 manually annotated images of all speakers in various poses, expressions and face occlusions, which constitute a training dataset for the ESR

detector. The database was recorded in an office environment using Microsoft Kinect sensor and Genius lavalier microphone.

Because of uncertainty of the stereo vision reconstruction in Kinect, there exist points in space, for which the depth is ambiguous and cannot be inferred without further assumptions about the observed scene. In such cases (e.g. inside of an opened mouth or around the nose), the Kinect device returns zero values. In order not to have skewed results, we therefore reconstruct all missing values in the depth maps by using nearest neighbor interpolation. We then remove the mean and clamp the depth values to the range $[-30, 30]$ in order to remove occasional large spikes manifesting when the background is partially visible.

For both video and depth streams, the average pixel value of the whole utterance is subtracted from each ROI to partially remove differences in light conditions between sequences recorded in different time. ROI images are also whitened to remove correlations between adjacent pixels. For the audio, we down-sample the original 44.1 kHz signal to 16 kHz before parametrization by MFCC.

5 Experiments

In order to reduce the effect of overfitting, we employ the cross validation strategy in all our experiments. The database of 54 speakers is split in a $43 : 11$ ratio in 5 different combinations[1]. We trained the ESR detector and all visual features separately for each training/testing split. All of the reported results are the average word recognition accuracies achieved over the five different splits. We used the Spearmint [18] library for Bayesian optimization and HTK toolkit [15] as implementation of Hidden Markov Models.

We compare the autoencoder (AE) features with features extracted using 2D DCT and PCA. The DCT coefficients are sorted according to their average energy achieved on the training set. The features are evaluated in three settings: static, static+delta (Δ), and dynamic linear discriminant analysis (LDA) [1]. In case of static and delta features, we exhaustively search for the optimal number of DCT and PCA features by maximizing the classification score. For DCT and PCA the respective optimal dimensions were 22 and 28 for video, and 16 and 14 for depth. The number of AE features is fixed to 144. In case of LDA, we reduce feature vector of each frame to 33 coefficients, concatenate $(2K + 1)$ neighboring frames into a single hyper-vector, and then reduce its dimension using LDA. We set $K = 5$ as an empirically found optimum between performance and complexity (LDA-K5). We use phonemes as class labels for the LDA. As a final step we perform feature mean subtraction for each utterance, in order to increase robustness against between-speaker variation of the visual features.

Table 1 presents achieved recognition accuracies of the considered features. The recognition was performed using a whole word 14 state HMM. As can be seen, AE features outperform both DCT and PCA in all three settings. However, the difference is smaller for LDA case. This is probably caused by violating the assumption of shared covariance matrices between all phoneme classes. In

[1] One testing group contains only 10 speakers.

Table 1. Word accuracy [%] for visual features individually

	Video			Depth		
	Static	Δ	LDA-K5	Static	Δ	LDA-K5
DCT	63.5	71.5	76.6	56.5	59.3	71.2
PCA	59.0	68.4	77.3	63.1	68.3	72.0
AE	67.7	75.4	78.2	64.0	68.3	75.4

order to improve the results of dynamic LDA, we therefore selected subset of the AE features according to their variance and then whitened the reduced features before frame concatenation. Note that for DCT and PCA only basic dimension reduction is needed. The recognition accuracies and the optimal vector dimensions also suggest that the depth stream contain less useful information than video. However, as we shall see next, the information contained in the depth stream is to some extent complementary.

Table 2. Word accuracy [%] for combinations of video and depth features

	LDA-K5		LDA-K5
DCT-DCT	81.6	AE-DCT	84.3
DCT-PCA	81.0	AE-PCA	85.0
AE-AE	85.9	DCT-AE	83.8

Table 2 shows results achieved for selected feature combinations. The features were combined by a multi-stream HMM using 0.6 : 0.4 weight ratio (video:depth). As can be seen, the recognition accuracy was improved when incorporating both modalities via multi-stream HMM as compared to single-source models. This holds for all pairs of features, suggesting that the depth-based features are complementary to video-based features. Similarly to previous experiment, the best result was achieved when the features were extracted by autoencoder from both video and depth streams. The absolute increase of accuracy for AE when incorporating depth was 7.7 %, which corresponds to 35 % relative improvement of word error rate (WER).

We also evaluate the AE features with incorporated depth information in simulated noisy environment. Since our database was recorded in a relatively quiet environment, babble noise from the NOISEX [16] database was added to the clean audio artificially using various signal-to-noise ratios (SNR). For audio and video feature fusion, the weight ratio was set to 0.5:0.5. When combining all three modalities the weights were 0.5:0.3:0.2 for audio, video, and depth, respectively. We also compare the achieved results with Multi-band Spectral Subtraction algorithm [17], a popular method for audio enhancement. The results are presented in Fig. 2. The graph again confirms the benefit when incorporating depth data in the lip-reading task. As one can expect, the improvement of audio-visual fusion as compared to audio-only recognition is highest for low SNR.

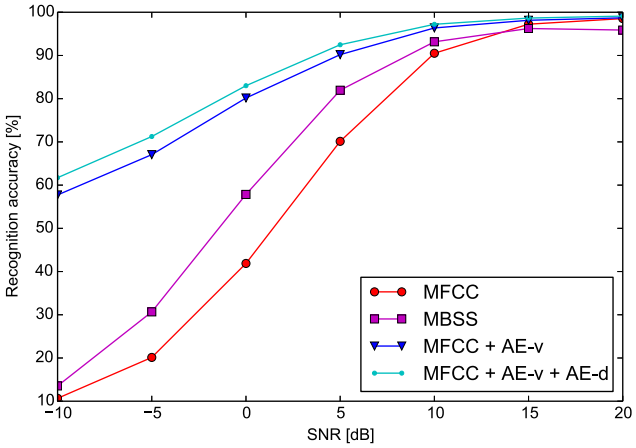

Fig. 2. Recognition accuracy in a noisy audio environment

The resulting scores are critically dependent on the individual stream weights, i.e. the results could be further improved by changing the weights dynamically depending on the SNR. Because of the spectral distortion, the MBSS algorithm is also beneficial especially for low SNRs.

6 Conclusions

We have evaluated the benefit of incorporating visual and depth information in the task of isolated word recognition. Based on the experiments we can conclude that the information contained in the depth data displays complementary character to the information captured by the video channel as confirmed by the 35 % relative WER reduction. In order to extract features from both video and depth streams, we have proposed to use a single layer autoencoder. We have shown an improvement of our autoencoder features over traditional techniques such as DCT and PCA for both video and depth channels. Compared to DCT, our AE features achieved 4–8 % higher absolute accuracy, depending on the data source and inclusion of speech dynamics. A disadvantage of our autoencoder features is higher dimensionality and additional required processing.

So far, the video and depth autoencoder features were evaluated only in the task of isolated word recognition. Our conclusions should also be confirmed in continuous speech recognition with phoneme-based models. The results achieved by autoencoder features could be potentially improved by utilizing deep learning algorithms for both video and depth streams. Also, the deep neural network could be utilized in other ways, e.g. as a feature fusion method or speech dynamics enhancement instead of LDA.

Acknowledgments. This work was supported in part by the Student Grant Scheme (SGS) at Technical University of Liberec.

References

1. Potamianos, G., Neti, C., Gravier, G., Garg, A., Senior, A.W.: Recent Advances in the Automatic Recognition of Audiovisual Speech. Proc. of the IEEE 91(9), 1306–1326 (2003)
2. Goecke, R.: Current Trends in Joint Audio-Video Signal Processing: A Review. In: Proc. of the Eighth International Symposium on Signal Processing and Its Applications, pp. 70–73 (2005)
3. Liew, A.W.Ch., W.S.: Visual Speech Recognition: Lip Segmentation and Mapping. Information Science Reference – Imprint. IGI Publishing, New York (2009)
4. Lan, Y., Theobald, B.J., Harvey, R., Bowden, R.: Comparing Visual Features for Lipreading. In: Proc. AVSP, pp. 102–106 (2009)
5. Paleček, K., Chaloupka, J.: Audio-visual Speech Recognition in Noisy Audio Environments. In: 36th International Conference on Telecommunications and Signal Processing (TSP), pp. 484–487 (2013)
6. Goecke, R., Millar, J.B., Zelinovsky, A., Ribes, R.J.: Stereo Vision Lip-Tracking for Audio-Video Speech Processing. In: Proceedings of the 2001 IEEE International Conference on Acoustics, Speech, Signal Processing (2001)
7. Císař, P., Krňoul, Z., Železný, M.: 3D Lip-Tracking for Audio-Visual Speech Recognition in Real Applications. In: Proc. INTERSPEECH (2004)
8. Galatas, G., Potamianos, G., Makedon, F.: Audio-visual Speech Recognition Incorporating Facial Depth Information Captured by the Kinect. In: Proc. EUSIPCO, pp. 2714–2717 (2012)
9. Pei, Y., Kim, T.-K., Zha, H.: Unsupervised Random Forest Manifold Alignment for Lipreading. In: Proc. ICCV, pp. 129–136 (2013)
10. Bengio, Y.: Learning Deep Architectures for AI. Foundations and Trends in Machine Learning 2(1), 1–127 (2009)
11. Ngiam, J., Khosla, A., Kim, M., Nam, J., Lee, H., Ng, A.Y.: Multimodal Deep Learning. In: Proc. ICML, pp. 689–696 (2011)
12. Huang, J., Kingsbury, B.: Audio-visual Deep Learning for Noise Robust Speech Recognition. In: Proc. ICASSP, pp. 7596–7599 (2013)
13. Viola, P.A., Jones, M.J.: Robust Real-Time Face Detection. International Journal of Computer Vision 57, 137–154 (2004)
14. Cao, X., Wei, Y., Wen, F., Sun, J.: Face Alignment by Explicit Shape Regression. In: Proc. CVPR, pp. 2887–2894 (2012)
15. Steve, Y., Odel, J., Ollason, D., Valtchev, V., Woodland, P.: The HTK Book, version 2.1. Cambridge University, United Kingdom (1997)
16. Varga, A.P., Steeneken, H.J.M., Tomlinson, M., Jones, D.: The NOISEX-92 Study on the Effect of Additive Noise on Automatic Speech Recognition. Technical Report, DRA Speech Research Unit (1992)
17. Kamath, S., Loizou, P.: A Multi-band Spectral Subtraction Method for Enhancing Speech Corrupted by Colored Noise. In: Proc. ICASSP, pp. IV-4164 (2002)
18. Snoek, J., Larochelle, H., Adams, R.P.: Practical Bayesian Optimization of Machine Learning Algorithms. Advances in Neural Information Processing Systems 25, 2951–2959 (2012)