v0 - A Multi-view Auto Lip-reading
The Kip-reasder's (Progress Report)

Michael Shell
School of Electrical and
Computer Engineering
Georgia Institute of Technology
Atlanta, Georgia 30332–0250
Email: http://www.michaelshell.org/contact.html

Homer Simpson
Twentieth Century Fox
Springfield, USA
Email: homer@thesimpsons.com

James Kirk
and Montgomery Scott
Starfleet Academy
San Francisco, California 96678–2391
Telephone: (800) 555–1212
Fax: (888) 555–1212

***Abstract*—The abstract goes here.**

I. Introduction

Lip-reading is a technique to understand speech by visually interpreting the movement of the lips, face and thought[1]. This technique is not limited to deaf or hear-of-hearing people but is also used by people which have a normal hearing process. A phanomena known as the McGurk effect[1] show this relation, where a change in interpretation of the same sound is changed with the visual input. Just as people use lip-reading for speech recognition it is also seen to have its application in artificial, where a higher accuracy can be obtained by combining the acoustic and visual information [2]. In artificial intelligent the combination of acoustic and visual information is know as audio-visual speech recognition (AVSR) and system where only the visual information is used in commonly known as automatic lip-readig (ALR) or visual speech recognition (VSR). ALR also have other promising applications, beside the combination with acoustic information, such as visual password, silent speech interface and forensic video analysis.

The main challenge in ALR is duo to a large variation in visual factors both from the recording such as changes in illumination and camera angle[3], but also from factors that is person specific such as mouth shape an visual pronunciation.

In order to address these challenges individually several experimental setup is proposed for ALR, such as: Speaker dependent (SD) or speaker independent (SI), single-view or multi-view. SD is the simplest setting where the personal variation from the speaker is removed since data from one speaker is used both for the training and evaluation. In SI the variation from the speaker such as mouth shape and visual pronunciation is included where unseen speakers is used for the evaluation. Single-view eliminate the change in camera angle where a multi-view setting include this dependency.

Previous work in ALR can in generally be grouped in to two, one with a classical approach and a more recent where deep neural networks is used.

In the classical approach the visual feature extraction is based on methods as component analysis, discrete wavelet transform, discrete cosine transform, active appearance model [5], local binary pattern [6], optical flow [7], Eigenlips [8], histograms of oriented gradients [9], internal motion histograms, motion boundary histograms and their mixed models [8, 10]. For multi-viewpoint lip-reading [11] adopt a minimum cross-pose variance analysis technique.[2]

Better performing systems is later obtained by the use of neural networks, where neural networks both have its application for both the feature extraction and the temporal correlation. In [2] a deep autoencoder for the feature extraction (Are there any temporal model in this architecture?). A long short-term memory (LSTM) is used in [4] to do a temporal correlation of the features. A state-of-the art performance is then obtained in [5], where a convolutional neural network (CNN) is used for the feature extraction together with a LSTM for the temporal correlation.

In this work we propose a combined solution between the classical approach and the neural network approach, by using techniques from the classical approach for the feature extraction in combination with a CNN and LSTM. We limit our scope to focus on speaker independent and multi-view setting, where we use the OuluVS2 database[6] to evaluate our design performance, in relation to the error rate of word or phrase classification. This task is related to the challenges given at the ACCV 2016 workshop, multi-view lip-reading/audio-visual challenge[3] (MLAC 2016).

II. Related Work

III. Background

A. Convolutional Neural Network

B. Long Short-Term Memory

C. Optical Flow

D. Key features

IV. Dataset

To train and evaluate our approach we use the public available OuluVS2 database. The recording environment used is illustrated in figure 1, where the multi-view setting can

[1] Wiki

[2] Taken from Multi-View Automatic

[3] http://ouluvs2.cse.oulu.fi/ACCVE.html

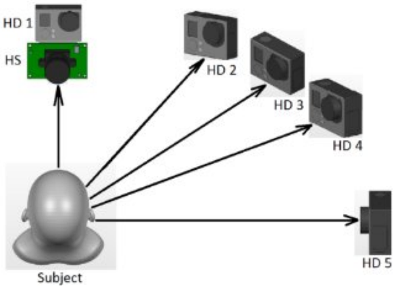

Fig. 1. Illustration of the multi-view setup used in the OuluVS2 recording, reference to figure???

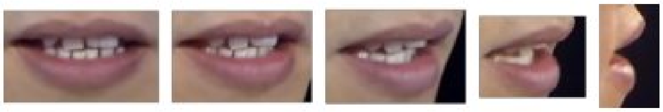

Fig. 2. Example of the preprocessed data with with the region of interest.

be seen. The setup make use of five cameras located at different angels in relation to the subject. For each recorded subject three different scenarios is performed, namely the pronunciation of: (1) a sequence of ten fixed digit sequences, (2) ten daily-use short English phrases and (3) five randomly selected TIMIT sentences. Examples of each of these can be seen in table I A total of 52 different test subject is used, where each subject is pronouncing each phrase three times.

A preprocessed version is available where the recording from different angles is synchronized and the region of interest (ROI) is both segmented, with the recording of interest, and cropped to only contain the mouth region. An example of the preprocessed data can be seen in figure 2.

V. Method

Alex-net + LSTM CNN-original + Bidirectional LSTM Optical flow + LSTM Key features + LSTM

VI. Result

VII. Conclusion

The conclusion goes here.

References

[1] H. McGurk and J. MacDonald, "Hearing lips and seeing voices," vol. 264, pp. 746–748, Dec. 1976.

[2] J. Ngiam, A. Khosla, M. Kim, J. Nam, H. Lee, and A. Y. Ng, "Multimodal Deep Learning," *Proceedings of The 28th International Conference on Machine Learning (ICML)*, pp. 689–696, 2011.

[3] G. Potamianos and C. Neti, "Audio-visual speech recognition in challenging environments." *Interspeech*, pp. 1293–1296, 2003. [Online]. Available: http://www.tsi.enst.fr/chollet/Biblio/Articles/Domaines/BIOMET/AudioVisual/Old/EURO03_CHALLENGING.pdf

[4] M. Wand, J. Koutnik, and J. Schmidhuber, "Lipreading with long short-term memory," in *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, mar 2016, pp. 6115–6119. [Online]. Available: http://ieeexplore.ieee.org/document/7472852/

[5] D. Lee, J. Lee, and K.-e. Kim, "Multi-View Automatic Lip-Reading using Neural Network," pp. 1–14.

[6] I. Anina, Z. Zhou, G. Zhao, and M. Pietikainen, "OuluVS2: A multi-view audiovisual database for non-rigid mouth motion analysis," *2015 11th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition, FG 2015*, 2015.

(1) Digits	”1 7 3 5 1 6 2 6 6 7” ”4 0 2 9 1 8 5 9 0 4”
(2) Phrases	”Thank you” ”Have a good time”
(3) TIMIT	”Chocolate and roses never fail as a romantic gift”

TABLE I
Examples of the three different scenarios used