# Robust Multi-Modal Speech Recognition in Two Languages Utilizing Video and Distance Information from the Kinect

Georgios Galatas[1,2], Gerasimos Potamianos[3,2], and Fillia Makedon[1]

[1] Heracleia Human Centered Computing Lab,
Computer Science and Engineering Dept.,
University of Texas at Arlington, USA
[2] Institute of Informatics and Telecommunications,
NCSR "Demokritos", Athens, Greece
[3] Dept. of Computer and Communication Engineering,
University of Thessaly, Volos, Greece
georgios.galatas@mavs.uta.edu, gpotam@ieee.org,
makedon@uta.edu

**Abstract.** We investigate the performance of our audio-visual speech recognition system in both English and Greek under the influence of audio noise. We present the architecture of our recently built system that utilizes information from three streams including 3-D distance measurements. The feature extraction approach used is based on the discrete cosine transform and linear discriminant analysis. Data fusion is employed using state-synchronous hidden Markov models. Our experiments were conducted on our recently collected database under a multi-speaker configuration and resulted in higher performance and robustness in comparison to an audio-only recognizer.

**Keywords:** Audio-visual automatic speech recognition, multi-sensory fusion, languages, linear discriminant analysis, depth information, Microsoft Kinect.

## 1 Introduction

Speech is the most natural form of communication for humans, and therefore automatic speech recognition (ASR) is one of the most intuitive forms of human-computer interaction (HCI). To improve ASR accuracy and robustness to noise, incorporation of visual information in conjunction with audio has been shown to be beneficial [1, 2]. However, in most research studies, such information is obtained from traditional planar video, thus not utilizing 3D visual speech articulation information. To alleviate this shortcoming, only a handful of efforts have appeared employing multiple or stereo cameras to capture the speaker's face [3-5], with an increase though in hardware cost and software complexity. We have recently proposed an alternative to such approach, by aiming to capture 3D visual speech information from the depth sensor of the novel Kinect device that operates based on the structured light method [6]. That work however considered audio-visual speech recognition (AVASR) in English only [7].

M. Kurosu (Ed.): Human-Computer Interaction, Part IV, HCII 2013, LNCS 8007, pp. 43–48, 2013.

In this paper, we extend our previous work to consider AVASR in Greek, deviating from the traditional AVASR literature paradigm that considers one language only. Our system has been tested using two different languages, English and Greek, in a tri-stream multimodal fusion approach to ASR, where audio, planar video and distance information are combined for a small-vocabulary recognition task in order to keep data collection at a manageable level. Our experiments demonstrate consistent benefits when using the additional modalities to the performance and robustness for the ASR task across the two languages considered.

The design and experimentation using our system is presented in the next sections as follows: Initially, the system architecture is presented in Section 2 with details about the visual feature extraction and fusion. The experimental setup and results are discussed in Section 3 and our conclusions are presented in Section 4.

## 2 Description of the System Architecture

The input streams used by our system are audio, planar video and the distance information stream captured by the Kinect. The audio stream was captured using a Zoom H4 external voice recorder exhibiting good directionality and frequency response at 16-bit, 44.1kHz, PCM format. The planar video (24-bit color, VGA resolution) and the distance information (11-bit, VGA resolution) were both captured using the Kinect. The system architecture is shown in figure 1, and the various modules of the system are described in more detail in the following paragraphs.

### 2.1 Visual Front-End

The visual front-end is responsible for detecting and tracking the mouth region of interest (ROI) from each video frame. A nested setup using 2 Viola-Jones detectors [8] is used to detect the face and mouth of the speaker respectively. The nested implementation minimizes the number of false mouth detections by only searching for a mouth if the face of the speaker has already been detected. In addition, the coordinates of the mouth bounding box are smoothed by a median filter in order to minimize abrupt movements due to false detections. The detected mouth ROI coordinates from the video stream are also used for extracting the mouth region from the distance information stream. Finally, the size of both ROIs is normalized to 64x64 pixels.

### 2.2 Feature Extraction and Selection

The next step is to extract meaningful features from all 3 streams. For the audio stream, the well known Mel frequency cepstral coefficients (MFCCs) are extracted using the "Hidden Markov Model Toolkit" (HTK) [9], reaching a dimensionality of 39 (including first and second derivatives). For the video and distance streams, the coefficients of the 2-D discrete cosine transform (DCT) are extracted and interpolated to 100 Hz to match the rate of the audio features. Following is the feature selection step, which is comprised of 2 parts. Initially, the 45 highest energy coefficients of the upper left corner of each DCT image are selected as those with the highest information content. Subsequently, linear discriminant analysis (LDA) is applied to the features and those corresponding to the highest eigenvalues are selected as the most informative ones. Finally, the first and second derivatives are appended to the final feature vector in order to capture the dynamics of speech. The final feature dimensionality is 21 for each of the video and distance streams.

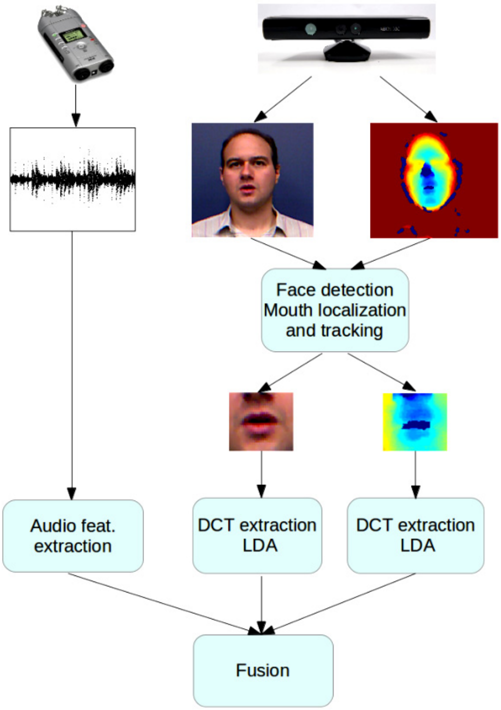

**Fig. 1.** Overview of the system modules

### 2.3 Data Fusion and Modeling

Hidden Markov models (HMMs) are the most commonly used classifiers for modeling speech. In our experiments we utilized state-synchronous multi-stream HMMs in order to effectively fuse the data from all 3 streams.

$$\Pr[o_t^{AVD} \mid c] = \prod_{s\in\{A,V,D\}} [\sum_{k=1}^{Ksc} \omega_{sck} N_{d_s}(o_t^{(s)}; m_{sck}, s_{sck})]^{\lambda_{sct}} \qquad (1)$$

This type of model realizes a decision-fusion approach, by computing the state emission (class conditional) probability as a product of the observation likelihoods of every stream, raised to a specific exponent $\lambda$, as shown in eq. 1. This exponent is bound to the reliability of the stream itself and defines the contribution of each stream. $o_t^{AVD}$ denotes the tri-modal observation vector $o_t^{AVD} = \{o_t^A, o_t^V, o_t^D\}$, $s$ is one of the three streams, $c$ denotes the HMM state and $t$ is the time (frame) of the utterance. The HMMs used in our work have a 3-state left-to-right topology, modeling tri-phones with 16 Gaussian mixtures per stream and state. HTK patched with HTS [10] were used for training and testing using the aforementioned models.

## 3 Experimental Results and Discussion

To support this work, we have captured our own database, the bilingual audio-visual corpus with depth information or BAVCD [11], that includes audio, planar video, and distance measurements using a voice recorder, the Kinect, and an HD camera. The corpus contains data from 15 speakers for the English part and 6 speakers for the Greek part, uttering connected digit strings.

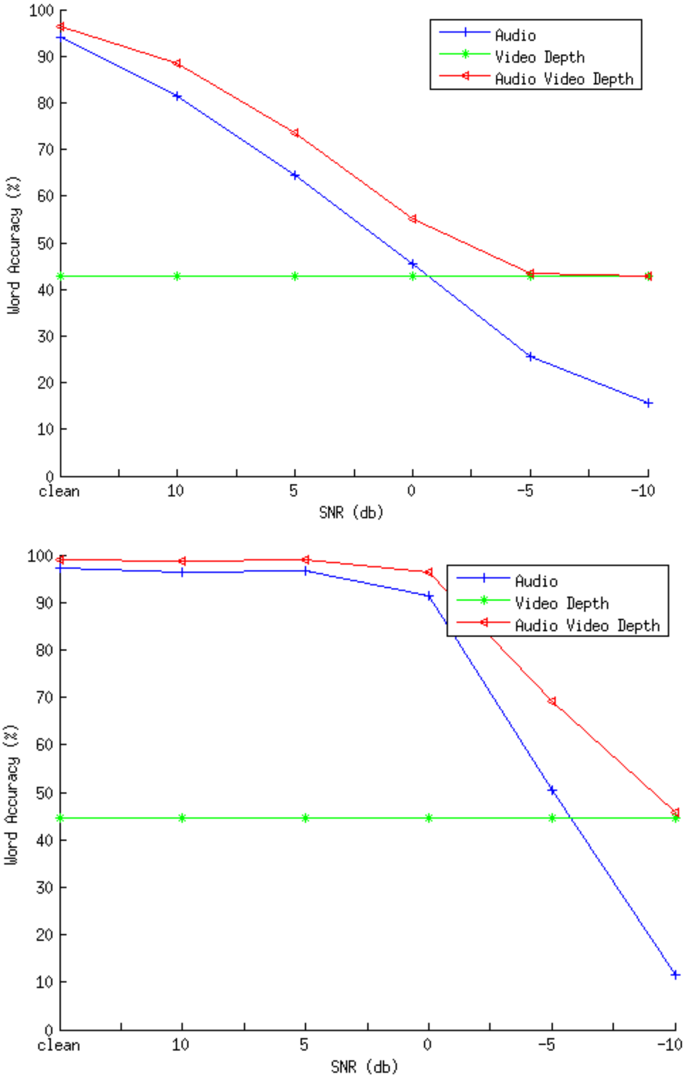

**Fig. 2.** Word accuracy results for English (top) and Greek (bottom) under various SNR levels

Our system was extensively tested in a multi-speaker setup under a variety of babble noise levels from the Noisex-92 database [12] in order to simulate a realistic smart home environment. Furthermore, training was conducted on clean speech, simulating the mismatch between the training and testing conditions. The results for each language are shown in figure 2. The best performance was achieved in the lack of noise when all 3 streams were utilized by the system. Under these conditions, the word accuracy for the Greek part was 99.02% and for the English part 96.32%. The performance for lip-reading without using audio was consistent for both parts and exhibited a 9.2% relative improvement using LDA. The overall system performance, degraded as the audio noise levels grew higher, but always remained higher than the performance of the individual streams, leading to a significant improvement under very noisy conditions e.g. 45.63% instead of 11.55% word accuracy for our system in comparison to an audio-only recognizer for a signal to noise ratio (SNR) of -10dB in Greek. The system exhibited better performance for the Greek language due to the smaller number of Greek speakers in the database in conjunction with the multi-speaker setup of the experiments.

## 4 Conclusions

In conclusion, we developed a novel speech recognition system that in addition to audio utilizes planar video and distance measurements captured by the Kinect and we tested its performance in both English and Greek. We have shown that our system exhibits high recognition rates in clean audio conditions but is also robust in noisy conditions, achieving significantly higher performance than an audio-only ASR system. Finally, our system's performance is consistent in both languages, constituting a reliable solution for speech recognition.

**Acknowledgments.** This material is based upon work supported by the National Science Foundation under Grants No. NSF-CNS 1035913, NSF-CNS 0923494. G. Potamianos would like to acknowledge partial support from the European Commission through FP7-PEOPLE-2009-RG-247948 grant AVISPIRE.

## References

1. Iwano, K., Tamura, S., Furui, S.: Bimodal speech recognition using lip movement measured by optical-flow analysis. In: Proc. HSC, pp. 187–190 (2001)
2. Nakamura, S., Ito, H., Shikano, K.: Stream weight optimization of speech and lip image sequence for audio-visual speech recognition. In: Proc. ICSLP, vol. 3, pp. 20–24 (2000)
3. Goecke, R., Millar, B.: The audio-video Australian English speech data corpus AVOZES. In: Proc. ICSLP, vol. 3, pp. 2525–2528 (2004)
4. Vorwerk, A., Wang, X., Kolossa, D., Zeiler, S., Orglmeister, R.: WAPUSK20 – a database for robust audiovisual speech recognition. In: Proc. LREC (2010)
5. Ortega, A., Sukno, F., Lleida, E., Frangi, A., Miguel, A., Buera, L., Zacur, E.: AV@CAR: A Spanish multichannel multimodal corpus for in-vehicle automatic audio-visual speech recognition. In: Proc. LREC., vol. 3, pp. 763–767 (2004)

6. The Primesensor Reference Design, `http://www.primesensor.com`
7. 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)
8. Bishop, C.M.: Pattern Recognition and Machine Learning. Springer (2006)
9. Young, S.J., Kershaw, D., Odell, J., Ollason, D., Valtchev, V., Woodland, P.: The HTK Book version 3.4. Cambridge University Press (2006)
10. The HMM-based speech synthesis system (HTS), `http://hts.sp.nitech.ac.jp`
11. Galatas, G., Potamianos, G., Kosmopoulos, D., Mcmurrough, C., Makedon, F.: Bilingual corpus for AVASR using multiple sensors and depth information. In: Proc. AVSP, pp. 103–106 (2011)
12. Varga, A., Steeneken, H.: Assessment for automatic speech recognition: Noisex-92. A database and an experiment to study the effect of additive noise on speech recognition systems. Speech Communication 12(3), 247–251 (1993)