SUTAV: A Turkish Audio-Visual Database

Ibrahim Saygin Topkaya and Hakan Erdogan

Vision and Pattern Analysis Laboratory
Sabanci University - Faculty of Engineering and Natural Sciences
Tuzla, 34956, Istanbul, Turkey
{isaygint,haerdogan}@sabanciuniv.edu

Abstract

This paper contains information about the "Sabanci University Turkish Audio-Visual (SUTAV)" database. The main aim of collecting SUTAV database was to obtain a large audio-visual collection of spoken words, numbers and sentences in Turkish language. The database was collected between 2006 and 2010 during "Novel approaches in audio-visual speech recognition" project which is funded by The Scientific and Technological Research Council of Turkey (TUBITAK). First part of the database contains a large corpus of Turkish language and contains standart quality videos. The second part is relatively small compared to the first one and contains recordings of spoken digits in high quality videos. Although the main aim to collect SUTAV database was to obtain a database for audio-visual speech recognition applications, it also contains useful data that can be used in other kinds of multimodal research like biometric security and person verification. The paper presents information about the data collection process and the the spoken content. It also contains a sample evaluation protocol and recognition results that are obtained with a small portion of the database.

Keywords: Audio-Visual Speech Recognition, Turkish Audio-Visual Database, Multi-Modal Data Collection

1. Introduction

With the increasing power in computing, there has been greater room for applications of multiple modalities in human-computer interaction (HCI). Audio and visual modalities are considered as major modalities of HCI and probably the most natural way of interaction if interaction between people are considered. The conventional audio based applications like speech recognition and identity verification has begun to include visual modalities thus evolving into audio-visual applications (Dupont and Luettin, 2000; Acheroy et al., 1996).

During our research on audio-visual speech recognition within the "Novel Approaches in Audio Visual Speech Recognition" research project, a requirement of an audio-visual database of Turkish language had arisen; thus the idea of collecting this data has been born. Although the primary aim of the collection process was to create a Turkish database to be used in audio-visual speech recognition, the content of the database has been set broader to include data to be used in other types of multi-modal research and development systems like personal identity verification systems.

In this paper we present the SUTAV (Sabanci University Turkish Audio-Visual) database; which consists of a large audio-visual collection of spoken sentences in Turkish language. The database includes different sets of videos recorded in a long time span and different people and recording conditions and subsets of different videos may be used for specific research purposes. This paper is organized as follows; in the next two sections we give information about the data collection process and the types of spoken examples that exist in the database. In the following section we present some recognition results on a subset of the database where an example evaluation protocol is implicitly included. In the final section we give information about the distribution of the database.

2. Data Collection and Format

SUTAV database had been collected between 2006 and 2010 within a four year long research project. The main motivation behind SUTAV was obtaining a Turkish database to be used in audio-visual speech recognition and person identification research where databases for French (Pigeon and Vandendorpe, 1997) and English (Messer et al., 1999) languages are commonly available. The collection process had been held as an undergraduate project course so that recorded subjects are gathered and organized as undergraduate students, which had allowed collecting data from a vast number of subjects.

The major part of the database is recorded as standard definition (SD) videos using Sony DCR HC23 camera connected to an internal FireWire interface and Rode directional microphone connected to an external M-Audio Fast Track Pro Audio/MIDI interface. These recordings have been acquired using a personal computer and encoded as audio video interleave (DV Codec) files. The videos have 720x576 pixels resolution and 25 fps frame rate having audio streams at 44.1 khz sampling rate. The audio and visual data have been encoded together into the video file simultaneously during recording process.

For the recordings that are gathered during 2010, Sony HD camera had been used. The videos that are obtained for this subset are high definition (HD) quality, where audio and visual data are both acquired with the camera only and encoded on the camera as advanced video coding high definition (AVCHD/MTS) files. The videos have 1920x1080 pixels resolution and 50 fps frame rate having 5 channels audio streams at 48 khz sampling rate. Since HD recordings were only performed during the last few months of the process the number of recordings in HD quality are relatively low compared to the SD ones.

Since the recorded data per file is relatively short, the subjects did not read the recordings from any paper or screen, and were instructed to direct their gaze to the camera. In the

first three sets (explained in next section) of the recording process the subjects were not specially illuminated apart from the ceiling illumination in the room, however in the last three sets of SD recordings and HD recordings the subjects' face were illuminated from the camera direction with halogen lamp. The room was not acoustically designed.

3. Database Content

The content of SUTAV varies between SD and HD recordings.

3.1. Content of SD Recordings

SD recordings consist of a large number of recordings in different sessions. In these sessions the subjects speak:

1. Their own names
2. A few previous subjects' names
3. Counting digits between zero and nine
4. Random digit quartets
   e.g. "Dokuz-Bes-Dort-Bir (Nine-Five-Four-One in Turkish)"
5. Random names of four cities in Turkey
   e.g. "Kadikoy-Ankara-Eskisehir-Bergama"
6. Random phonetically rich Turkish sentences
   e.g. "Sabirsizlikla kardesinden gelecek telefonu bekliyordu"

Random numbers, names and sentences are selected from a large pool of pre-determined sets and differ periodically from subject to subject. There are a total of 100 sentences spoken, 15 of which are taken from TURTEL database (Yapanel et al., 2001) and the rest are selected from various sources to cover the phonetic properties of Turkish language. This allows to have a large corpus spoken by different people, however also some different people speaking the same sets allowing comparison of techniques on different people with the same content. Counted digits consist in a Turkish spoken digit database which is comparable with M2VTS (Pigeon and Vandendorpe, 1997) and XM2VTS (Messer et al., 1999) which are available in French and English respectively.
SD recordings are recorded in different sessions and different number of total videos for each subject. The recordings are organized in recording terms (i.e. semesters) and recording sessions (i.e. tapes). Each subject's recording including all sessions is done within a specific term. The terms and contents of the SD recordings are as follows, as well as summarized in a table in Table 1:

2006-2007 (1): In this set 21 female and 23 male subjects are recorded. The recordings begin with a rotated head shot of the subject as in Figure 1, followed by five videos where subject says their name, five videos where subject says the name of previous five subjects, three videos where subject counts from zero to nine, ten videos of random sentences, ten videos of random digit quartets and four videos of random city quartets where in total thirty-eight videos had been recorded. Each subject had been recorded in two sessions where in each session the spoken content is the same.

(a) (b) (c)

Figure 1: Sample frames from a recording where the subject rotates her head.

2007-2008 (1): In this set 24 female and 51 male subjects are recorded. The content of recordings is the same.

2007-2008 (2): In this set 20 female and 10 male subjects are recorded. The content of recordings is the same. However a third session for each subject is also recorded where colored dot markings are put on the lips of the subject which may help lip tracking applications.

2008-2009 (1): In this set 19 female and 36 male subjects are recorded. The content of recordings is the same and held in three sessions, however without applying any dot markings on the lips as in Figure 2.

2008-2009 (2): In this set 22 female and 41 male subjects are recorded. The recordings begin with a rotated head shot of the subject, followed by five videos where each subject says their name, five videos where subject says the name of previous five subjects, one video where subject counts from zero to nine, five videos of random sentences, five videos of random digit quartets and five videos of random city quartets where in total twenty-seven videos had been recorded. Each subject had been recorded in three sessions where in each session the spoken content is the same.

2009-2010 (1): In this set 18 female and 17 male subjects are recorded. The content of recordings is the same as the previous set.

The SD recordings contain different types of videos suitable for different multi-modal applications. For example, spoken sentences and city names are suitable for continuous speech recognition or simple word sequence recognition applications. Digit recordings can also be used for word sequence recognition or security applications (like in a scenario where the digits are taken as passphrases). The videos where the subjects speak their own names and other subjects' names can be used on biometric applications where example sets of true identities (subjects saying their names) and false identities (subjects saying other subjects' names) are available.

3.2. Content of HD Recordings

In the last term of the SUTAV data collection process, HD quality videos are recorded. The content of HD quality

Number of Recordings	2006 - 2007 (1)	2007 - 2008 (1)	2007 - 2008 (2)	2008 - 2009 (1)	2008 - 2009 (2)	2009 - 2010 (1)
Rotated head shot	1	1	1	1	1	1
Subject's Name	5	5	5	5	5	5
Other Subjects' Names	5	5	5	5	5	5
Count 0...9	3	3	3	3	1	1
Sentences	10	10	10	10	5	5
Number Quartets	10	10	10	10	5	5
City Name Quartets	4	4	4	4	5	5
Female Subjects	21	24	20	19	22	18
Male Subjects	23	51	10	36	41	17
Number of Sessions	2	2	3 [1]	3	3	3

Table 1: Overall information about the content of SD recordings.
[1] Last session is recorded with lip markers as shown in Figure 2.

videos are narrower compared to SD ones and contain the recording of subjects saying five number quartets. In total 141 subjects were recorded where each subject has only one session of recording. This data had been collected to perform real-time digit recognition experiments, so the collection process is limited to digits only and since the whole data is collected to be used as a training set for the real-time experiments, different sessions were not managed.

4. An Example Evaluation Protocol and Recognition Results

To test the audio-visual speech recognition (Dupont and Luettin, 2000) results on the collected data as compared to the M2VTS database (Pigeon and Vandendorpe, 1997), we have used a small subset of the data where the subjects count digits from zero to nine. The motivation behind selecting this subset is to have a constant word sequence across all videos like the M2VTS database. We have selected videos recorded in three terms (2008-2009 (1), 2008-2009 (2) and 2009-2010 (1)) all of which have three sessions. The results are obtained by selecting first two sessions as training set and the last session as testing set. In a previous work (Topkaya et al., 2011) we had applied a novel proposed approach to M2VTS, and here we apply the same approach to the selected subset of SUTAV database. A short summary of the applied approach is explained in this section.

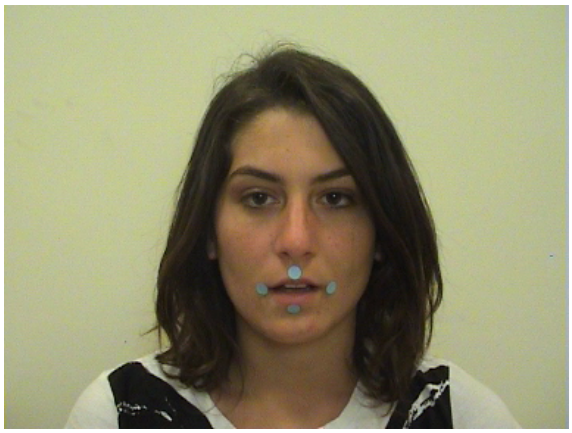

Figure 2: Sample frame from a recording showing the subject with a colored dot marking on the lips.

We benefit from hidden Markov models (HMM) (Rabiner and Juang, 1986) to perform speech recognition. First we train audio only HMMs on clean audio data on training set and use these to extract frame by frame alignment information for every video on the training set. Then we use data for each frame in training set to train first-level discriminative (named "tandem" (Hermansky et al., 2000)) classifiers where each phoneme corresponds to one class. Using these classifiers we extract posterior probabilities of each class (i.e. phoneme) for every frame on the whole set for audio and visual data separately. This results in frame by frame additional streams for the whole data set which contains posterior probabilities for each phoneme.

We use two kinds of classifiers; Support Vector Machines (Burges, 1998) and Neural Networks (Haykin, 1999) and use outputs of these classifiers as two additional streams as well as audio and visual streams in a multi-stream hidden Markov model (MSHMM). We refer to this stream combination as "visual tandem" approach. We also use two kinds of classifier combiners (Erdogan and Sen, 2010) and find an optimal weighted combination of audio and visual tandem classifier outputs and use these combinations as two additional streams to audio and visual streams. We refer to this second kind of stream combination as "'tandem fusion" approach.

To test the data, audio noise has been added in different SNR levels. Then second tape in the training set is separated as validation set and for each SNR level optimal weights for MSHMM streams are found. These weights are used for the final recognition result on the MSHMMs trained on the whole training set and applied on the test set. The results for different SNR levels and different kinds of stream combinations as well as optimal weights (inferred from validation set) and applied on test set are presented in

SNR	Only Audio	Only Visual	Audio Visual	Visual Tandem	Tandem Fusion	Audio Weight	Visual Weight
Clean	92.32	29.12	92.32	92.32	92.32	1	0
20	92.96	29.12	92.96	92.96	92.96	1	0
15	91.2	29.12	91.2	91.2	91.2	1	0
10	83.2	29.12	85.44	84.96	87.36	0.9	0.1
5	57.92	29.12	67.2	68	59.28	0.7	0.3
0	25.12	29.12	36.48	37.84	41.76	0.4	0.6
-5	13.04	29.12	28.56	34.24	38.16	0.1	0.9
-10	4.56	29.12	26.16	32.4	35.12	0.1	0.9
-15	0.32	29.12	29.12	34.56	38.16	0	1
-20	0.64	29.12	29.12	34.56	38.64	0	1

Table 2: Recognition accuracies on SUTAV database obtained with tandem fusion method.

Table 2.

The results show that under high noise level (i.e. low SNR) the contribution of visual data to the recognition accuracy increases, which is an indicator of benefits of using multi-modal approach in recognition. Columns with tandem streams contain tandem classifiers, that are employed before hidden Markov modeling and outputs of those classifiers are fed into hidden Markov models as observation vectors. Visual Tandem column holds tandem classifiers only for visual data and Tandem Fusion column holds tandem classifiers for both audio and visual data and combines them with suitable classifier fusion systems. It can be seen that employing tandem classifiers increases recognition accuracy even more.

This example experimental setup may give a clue about future research using the database. Multiple sessions of recordings allow researchers to handle data in smaller subsets for training, validation and testing.

5. Distribution of SUTAV Database

To obtain the whole database or a subset of it, please contact the authors. The distribution includes the recorded videos as well as content of spoken words and sentences per video, organized in terms and sessions as described in Section 3.

6. Acknowledgements

The database has been collected within the "Novel Approaches in Audio Visual Speech Recognition" research project (code 107E015) funded by The Scientific and Technological Research Council of Turkey (TUBITAK) under the scientific and technological research support program (code 1001).

The authors are also grateful to Sabanci University undergraduate students who participated in the collection process of the database.

7. References

M. Acheroy, C. Beumier, J. Bign, G. Chollet, B. Duc, S. Fischer, D. Genoud, P. Lockwood, G. Maitre, S. Pigeon, I. Pitas, K. Sobottka, and L. Vandendorpe. 1996. Multi-modal person verification tools using speech and images.

Christopher J.C. Burges. 1998. A tutorial on support vector machines for pattern recognition. *Data Mining and Knowledge Discovery*, 2:121–167.

Stphane Dupont and Juergen Luettin. 2000. Audio-Visual Speech Modelling for Continuous Speech Recognition. *IEEE Transactions on Multimedia*. to appear.

H. Erdogan and M.U. Sen. 2010. A unifying framework for learning the linear combiners for classifier ensembles. In *Pattern Recognition (ICPR), 2010 20th International Conference on*, pages 2985 –2988, aug.

Simon Haykin. 1999. *Neural Networks: A Comprehensive Foundation*. Prentice Hall.

Hynek Hermansky, Daniel P. W. Ellis, and Sangita Sharma. 2000. Tandem connectionist feature extraction for conventional hmm systems. In *ICASSP*.

K. Messer, J. Matas, J. Kittler, J. Lttin, and G. Maitre. 1999. Xm2vtsdb: The extended m2vts database. In *In Second International Conference on Audio and Video-based Biometric Person Authentication*, pages 72–77.

Stephane Pigeon and Luc Vandendorpe. 1997. The m2vts multimodal face database (release 1.00). In Josef Bign, Grard Chollet, and Gunilla Borgefors, editors, *Audio- and Video-based Biometric Person Authentication*, volume 1206 of *Lecture Notes in Computer Science*, pages 403–409. Springer Berlin / Heidelberg. 10.1007/BFb0016021.

L. R. Rabiner and B. H. Juang. 1986. An introduction to hidden Markov models. *IEEE ASSP Magazine*, 3(1):4–16, Jan.

I.S. Topkaya, M.U. Sen, M.B. Yilmaz, and H. Erdogan. 2011. Improving speech recognition with audio-visual tandem classifiers and their fusions. In *Signal Processing and Communications Applications (SIU), 2011 IEEE 19th Conference on*, pages 407 –410, april.

U. Yapanel, T. Islam, M. U. Dogan, and H. Palaz. 2001. Turtel database technical report. Technical report, TUBITAK-UEKAE.