BUILDING A DATA CORPUS FOR AUDIO-VISUAL SPEECH RECOGNITION

Alin G. Chiţu and Leon J.M. Rothkrantz
Man-Machine Interaction Group
Delft University of Technology
Mekelweg 4, 2628CD Delft,
The Netherlands
E-mail: {A.G.Chitu,L.J.M.Rothkrantz}@ewi.tudelft.nl

KEYWORDS
Audio-visual data corpus, lipreading, audio-visual speech recognition.

ABSTRACT

Data corpora are an important part of any audio-visual research. However, the time and effort needed to build a good dataset are very large. Therefore, we argue that the researchers should follow some general guidelines when building a corpus that guarantees that the resulted datasets have common properties. This will give the opportunity to compare the results of different approaches of different research groups even without sharing the same data corpus. In this paper we will formulate the set of guidelines that should always be taken into account when developing an audio-visual data corpus for bi-modal speech recognition. During the process we compare samples from different existing datasets, and give solutions for solving the drawbacks that these datasets suffer. In the end we give a complete list with all the properties of some of the most known data corpora.

INTRODUCTION

Data corpora are an important part of any audio-visual speech recognition research. Having a good data corpus, (i.e. well designed, capturing both general and also particular aspects of a certain process) might be of great help for the researchers in this field as it could greatly influence the research results. However, partly because the field is still young, or partly because the time and resources it takes to record a multi-modal data corpus can be overwhelming, the number of existing multi-modal data corpus is small compared to the number of uni-modal datasets. In order to evaluate the results of different approaches for a certain problem the data corpora should be shared between researchers or otherwise there should be some exact guidelines for building a corpus that all datasets should comply with. In the case when a data corpus is build with the intension to be made public, a greater level of reusability is required. In all cases, the first and probably the most important step in building a data corpus is to carefully state the targeted application(s) of the system that will be trained using the dataset. Currently the main applications of an audio-visual dataset are: audio-visual speech recognition (TULIPS1, AVletters, AVOZES, CUAVE, VidTIMIT, DAVID, IBM LVCSR and DUTAVSC), speaker detection, identity verification (VALID, M2VTS, XM2VTS, VidTIMIT and DAVID), user affective state recognition and talking heads generation.

In the current paper we will focus on the issues related to the audio-visual datasets built having the stated target speech recognition. From the point of view of speech recognition the common limitations that an audio-visual dataset has are:

- The recordings contain only a small number of respondents. This greatly reduces the generality of the results, since it generally generates highly under-trained systems. Hence not all the possible sources of variances are captured. The bias of such datasets comes from the fact that they are unbalanced with respect to gender, race and age of the respondents. Usually the number of respondents is a two digit number, with very few exceptions that use some 200-300 respondents. Even is these situations a good practice is to carefully record the speaker's data, such as age, gender, race, dialect, etc.
- The pool of utterances is usually very limited. The datasets usually contain only isolated words or digits or even only the letters of the alphabet rather than continuous speech. This induces a poor coverage of the set of phonemes and visemes in the language. Therefore, if continuous speech is targeted then the prompts used should always contain phonetically rich words and sentences. A good idea will be to search for the words that efficiently cover the possible combinations of phonemes in the language. This will help keeping the respondent's effort in reasonable limits. Moreover, phonetically rich speech will also better represent the co-articulatory effect in the language.
- The quality of the recordings is often very poor. This usually holds for the video data. It can be argued that for specific applications, such as speech recognition while driving, using dedicated databases (for instance AVICAR database; see Lee 2004) might better represent the specifics of the speech in this situation, but however the use of such dataset will be entirely restricted. We will show in the next sections what the main pitfalls are, and give exact workaround solutions.
- The datasets are not publicly available. Many datasets that are reported in scientific papers are not open to the public. This makes impossible the verification of the results of the different methods, and forces the researchers to build their own dataset.

One of the first datasets used for lipreading was TULIPS1 (Movellan 1995). This database was assembled in 1995 and consists of very short recordings of 12 subjects uttering the first 4 digits in English. Another very small dataset is AVletters (Matthews 1998). Later, other datasets were compiled which are larger and have a greater degree of usability, for instance ValidDB (Fox 2005), AVOZES (Goecke and Millar 2004), CUAVE (Patterson et al. 2002), VidTIMIT (Sanderson and Paliwal 2004) and IBM LVCSR.

In the following sections of the paper we will underline the main issues of the existing datasets for speech recognition with respect to audio and video quality in section 2 and 3, and with respect to language completeness in section 4. The different datasets will be compared during the process. In the comparison we introduce our own dataset DUTAVSC specially built for audio-visual speech recognition. Details about the DUTAVSC corpus can be found in the paper (Wojdeł et al. 2002).

AUDIO QUALITY

The complexity of audio data recording is much smaller than the one of the video recordings. The required hardware was developed long before speech recognition research was born. Therefore all datasets store the audio signal with sufficient high accuracy, namely using a sample rate of 22kHz to 48kHz and a sample size of 16bits. For comparison the audio CDs use a 44kHz sample rate with the same sample size per channel. Therefore the quality of the audio data should not be considered from the point of view of storage accuracy but from the perspective of recording conditions. It is interesting to know what the level of signal to noise ratio (SNR) is allowed during recordings. There could be two approaches here. Firstly, the database can be built with a very narrow application domain in mind such as speech recognition in the car. Also different versions for each possible situation can be recorded (for instance the dataset BANCA (Bailly-Baillire et al. 2003) built for identity verification has three versions for each of the three environments: controlled, degraded and adverse). However, the result is either a too dedicated dataset, or implies a large amount of work for building the dataset. Secondly, the dataset can be recorded in controlled, noise free environment and later on, following the necessities, the noise can be added to the recordings. The specific noise can be simulated or recorded in the required conditions and later superimposed on the clear audio data. An example of such database is NOISEX-92 (Varga and Steeneken 1993). This dataset contains white noise, pink noise, speech babble, factory noise, car interior noise, etc. For our dataset we used the second approach, and used white and pink noise to simulate a noisy environment.

VIDEO QUALITY

In the case of video data recording there are a larger number of important factors that control the success of the resulted data corpus. Hence, not only the environment, but also the equipment used for recording and other settings is actively influencing the final result. The environment where the recordings are made is very important since it can determine the illumination of the scene, and the background of the speakers. A large majority of the datasets were recorded indoors in controlled environment. In these cases the speaker's background was usually mono-chrome so that by using a "color keying" technique the speaker can be placed in different locations inducing in this way some degree of visual noise. In the case of dedicated datasets, as was shown in the previous section, the video data is also characteristic to the environment present at the location where the system is used.

Contrary to the audio case, the equipment used when recording plays a major role. Hence, while Tulips1 and AVletters datasets were compiled at the resolution of 100x75pixels and 80x60pixels, respectively, the newer datasets use much higher resolutions. For instance AVOZES, CUAVE uses 720x576pixels resolution. The same improvement in quality is also observed in the way the color information is sampled. The days of grayscale, 8bits per pixel images are long over. All datasets today save the color information using three channels each having 8 bits size. This is very important because the discriminatory information is highly degraded by converting to grayscale. The frame rate used is usually conforming to one of the color encoding systems used in broadcast television systems. Hence, by the place where the dataset was compiled we can have 24fps, 25fps, 30fps or 29.97fps recordings. Another important quality related property of the device used for recordings is the performance under changing illumination conditions. It is well know that available camcorders perform poorly under low illumination conditions. To alleviate this problem most video recording devices apply algorithms that increase the image intensity, however in chrome image information detriment. Therefore a good illumination is always required when recording. The light should be cast by all means at least uniformly on the scene, not to generate shadow patterns.

Another decision that needs to be made is where to focus, how large should be the scene? Should we define a region of interest (ROI), for instance only show the face of the speaker or maybe only show the lower half of the face, or show more background? Most of the datasets show however a passport like image of the speaker. We argue that defining a small ROI has many advantages and should be considered. Of course a much reduced ROI puts very high constraints on the performances of the video camera used and it might be argued that this is not the case in real life where the resulted system will be used. Recording only the mouth area as is done in the Tulips1 data set is clearly a very tough goal to achieve. However, by using a face detection algorithm combined with a face tracking algorithm we could automatically focus and zoom in on the face of the speaker. A small ROI facilitates acquiring a much greater detailed view of the interesting parts, while keeping the resolution of the frames in regular limits. To exemplify this in figure 1 is shown the area of the mouth as it is retrieved in some of the available datasets. The mouth area is manually clipped such that the bounding box touches the lower and upper lips and the left and right corners of the mouth. The frames were chosen such that to show approximately the same viseme, since is not possible to show exactly the same viseme in each picture. The resulted images were scaled up to a common size. During scaling some distortions appeared due to the fact that the obtained bounding boxes had different aspect ratios. The most visible distortions appeared in the case of the sample from AVOZES dataset. The smallest area reserved to the mouth is found in this example in the VidTIMIT dataset. The contrast of the images in this dataset is also quite poor. In general all datasets that have a large ROI, reserve a very small number of pixels to the mouth area which is however the main source of information for lipreading. In the DUTAVSC dataset, which was compiled in our group, we recorded only the lower part of the face which makes the mouth a central object in the image.

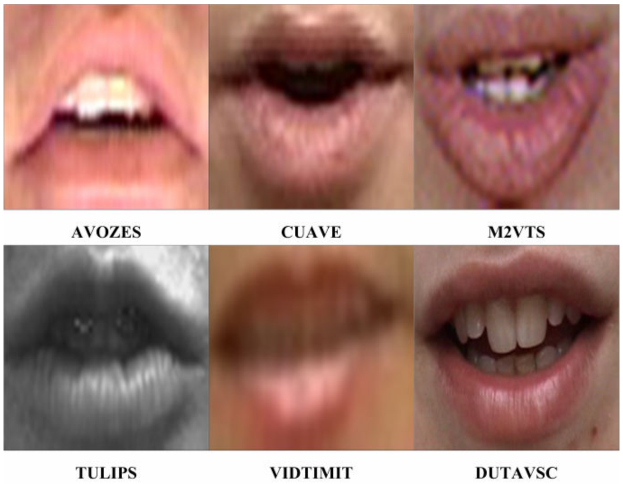

Figure 1: Quality of the ROI in Audio-Visual Data Corpora

Table 1 gives the sizes of the bounding boxes in all 6 samples. We see that the height in the case of DUTAVSC dataset is from 3.5 times to 5.5 times larger than in the case of the other datasets. The same pattern is seen when comparing the width of the bounding boxes, however a smaller difference from 2 times to 4.5 times larger is found.

Table 1: Sizes of the Bounding Boxes Surrounding the Mouth Area in 6 Different Datasets

Corpus	Width	Height
AVOZES	122	24
CUAVE	75	34
M2VTS	46	28
TULIPS1	76	37
VidTIMIT	53	25
DUTAVSC	225	133

During the recordings, attention should be paid to the way the respondents stand and move, especially when a small ROI is considered. If no automatic method is used for tracking the region of interest then the user should be very careful not to go out of the scene and not to move his head very much. Also during talking many speakers use their tongue to wet the lips. By doing so the mouth area will be covered making it impossible to recover correct information about what is being said. This will generate large amounts of noise in the resulted feature vectors. Random movement of the speaker head gives many problems to any method that attempts to extract movement information, for instance methods based on optical flow analysis. If this is the case, then the effect of head movement should be removed prior to feature extraction. The figure 2 shows some examples of broken clips recorded for the DUTAVSC dataset.

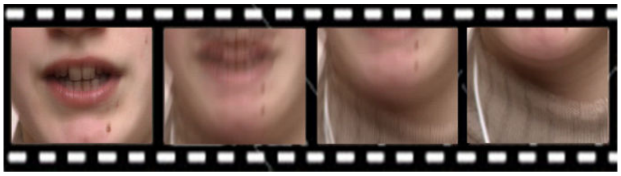

Figure 2: Faulty Clip Section from DUTAVSC Dataset

LANGUAGE QUALITY

As we said in the introduction, the quality of an audio-visual data corpus that has as target application speech recognition systems is measured by the degree of coverage of the phonemes and visemes of the targeted language. The number of phonemes differs from language to language as one may expect. For English there are 40 to 45 different phonemes depending on the dialect, for instance in Australian English there are 44 phonemes (Goecke 2005). In Dutch there are 42 different phonemes. In order to build a reliable system a good coverage of the phonemes is strongly required. For this purpose the utterances should contain phonetically rich words and sentences. The same goes for the visemes, which are the visual counterpart of phonemes. However the number of visemes is slightly smaller, for instance in English are 11 to 14 different visemes, while in Dutch 16. The words used should provide a good coverage of the combination of sounds, such as consonant vowel mixtures, so that all co-articulatory effects appear in a reasonable number of samples. Spelling samples should also be included in the dataset.

CONCLUSIONS

The quality of the data corpus used has a great impact on the results of the research initiative. For this reason, a good data corpus should be build following some strict rules that can guarantee the success of the final product. In this paper we emphasized the most important properties that a good data corpus should have, showed the main fallbacks of the current data corpora and give solutions that can alleviate the problems.

Table 2 lists the most well known data corpora to date and gives for each corpus the main characteristics with respect to their quality. The structure of the table is build such that to emphasize the quality of the datasets with respect to video, audio and language. In all cases only the gender of the speaker is recorded. We can see that our dataset scores quite reasonable.

Another aspect that was not covered until know by any of the data corpus is the frame rate at which a data set is recorded. All databases were recorder using low frame rates in conformity with the standard used at the location where the dataset was compiled. An interesting question is how is the frame rate of the video influencing the performances of the speech recognition system? When fusing the audio and video data for lipreading, a well know problem is the difference in the sample rate of the two data streams. Hence the audio stream is usually sampled at 100fps while the video stream will only provide 24-30fps. To solve this problem some up-sampling techniques need to be used. But what if the video data is recorded using a high speed camera so that the frame rate matches the audio frame rate? In order to tackle this question we plan to build such a database in the near future.

Table 2: Comparison Among Existing Corpora; Their Characteristics and Stated Purpose

Corpus	Language	Sessions	Respondents	Audio Quality	Video Quality	Language Quality	Stated purpose
TULIPS1	English	1	7male, 5female	11.1kHz, 8bits controlled audio	100x75, 8bit, 30fps mouth region	first 4 digits in English	small vocabulary isolated words recognition
AVletters	English	1	5male, 5female	22kHz, 16bits controlled audio	80x60, 8buts, 25fps mouth region	the English alphabet	spelling English alphabet
AVOZES	English	1	10male, 10female	48kHz, 16bits controlled audio	720x480, 24bits, 29.97fps entire face, stereo view	digits from ‘0’ to ‘9’ continuous speech application driven utterances	continuous speech recognition for Australian English
CUAVE	English	1	19male, 17female	44kHz, 16bits controlled audio	720x480, 24bits 29.970fps passport view	7,000 utterances connected and isolated digits	continuous speech recognition
Vid-TIMIT	English	3	24male, 19female	32kHz, 16bits controlled audio	512x384, 24bits, 25fps upper body	TIMIT corpus 10 sentences per person	automatic lipreading, face recognition
DAVID	English	12	132male, 126female (in 4 groups)	--	entire face, upper body, profile view multi corpora: controlled and degraded background, highlighted lips	vowel – consonants alternation, English digits	speech or person recognition
IBM LVCSR*	English	1	290 Unknown gender	22kHz, 16bits --	--	connected digits isolated words	audio-visual speech recognition
AVICAR	English	5	50male, 50female	48kHz, 16bits, 8channels 5 levels of noise car specific	4 cameras from different angles, passport view car environment	isolated digits, isolated letters, connected digits, TIMIT sentences	speech recognition in a car environment
DUTAVSC	Dutch	10-14	7male, 1female	48kHz, 16bits, controlled audio	384x288, 24bits, 25fps lower face view	spelling, connected digits, application driven utterances, POLYPHONE corpus**	audio-visual speech recognition, lipreading

* Not available to the public
** Data corpus for Dutch. Recordings are made over phone lines. More details can be found in (Damhuis et al. 1994)

REFERENCES

Bailly-Baillire, E.; Bengio, S.; Bimbot, F.; Hamouz, M.; Kittler, J.; Mariéthoz, J.; Matas, J.; Messer, K.; Popovici, V.; Porée, F.; Ruiz, B. and Thiran, J. 2003. “The BANCA Database and Evaluation Protocol” In *Proceedings of Audio and Video Based Biometric Person Authentication,* (Springer Berlin / Heidelberg, *2688*, pp. 625-638)

Damhuis, M.; Boogaart, T.; Veld, C. In't; Versteijlen, M.; Schelvis, W.; Bos, L. and Boves, L. 1994. "Creation and analysis of the Dutch polyphone corpus", In *ICSLP-1994,* (pp. 1803-1806)

Fox, N.A. 2005. “Audio and Video Based Person Identification”, PhD Thesis at *Department of Electronic and Electrical Engineering Faculty of Engineering and Architecture University College Dublin*

Goecke, R. and Millar, J. 2004. “The Audio-Video Australian English Speech Data Corpus AVOZES” *Proceedings of the 8th International Conference on Spoken Language Processing* (ICSLP2004, vol. III, 2525-2528)

Goecke, R. 2005. “Current Trends in Joint Audio-Video Signal Processing: A Review” In *Proceedings of the Eighth International Symposium on Signal Processing and Its Applications,* **(**August 28-31, pp. 70-73)

Lee, B.; Hasegawa-Johnson, M.; Goudeseune, C.; Kamdar, S.; Borys, S.; Liu, M. and Huang, T. 2004. “AVICAR: Audio-Visual Speech Corpus in a Car Environment” In *Proceedings of International Conference on Spoken Language Processing – INTERSPEECH2004*, (Jeju Island, Korea, October 4-8)

Matthews, I. 1998. “Features for Audio-Visual Speech Recognition” PhD thesis, School of Information Systems, University of East Anglia, October

Messer, K.; Matas, J. and Kittler, J. 1998. “Acquisition of a large database for biometric identity verification” In *BIOSIGNAL 98,* Vutium Press, (pp 70-72)

Movellan, J.R. 1995. “Visual Speech Recognition with Stochastic Networks” In *Advances in Neural Information Processing Systems, MIT Press*

Patterson, E.; Gurbuz, S.; Tufekci, Z. & Gowdy, J. 2002. “CUAVE: A New Audio-Visual Database for Multimodal Human-Computer Interface Research” In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*

Sanderson, C. and Paliwal K.K. 2004. “Identity Verification Using Speech and Face Information.” In *Digital Signal Processing* (vol. 14 nr. 5 pp. 449-480)

Wojdeł, J.C.; Wiggers, P. and Rothkrantz, L.J.M. 2002. “An audio-visual corpus for multimodal speech recognition in Dutch language” In *Proceedings of the International Conference on Spoken Language Processing (ICSLP2002)* (Denver CO, USA, September, pp. 1917-1920)

Varga, A. and Steeneken, H. 1993. “Assessment for automatic speech recognition II: NOISEX-92: a database and an experiment to study the effect of additive noise on speech recognition systems.” *Speech Communication*, (vol. 12, no. 3, pp. 247-251, July)

AUTHORS BIOGRAPHY

ALIN GAVRIL CHIȚU was born on November 8, 1978 in Bușteni, Romania. He graduated in 2001 at the Faculty of Mathematics and Computer Science at University of Bucharest, which is one of the top universities in Romania. In 2003 he received the MSc. degree in applied computer science at the same university. Starting September 2003 he joined the Risk and Environmental Master Program at Delft University of Technology, Delft, The Netherlands which he graduated with honors in August 2005. Since then he is pursuing his PhD degree in the Man-Machine Interaction Group, Mediamatics Department at Delft University of Technology under the supervision of Dr. Leon J.M. Rothkrantz. His main interest is in data fusion as the means to build robust and reliable systems, audio-visual speech recognition being one of the case studies. He is also interested in robust computer vision, machine learning and computer graphics.
Email: a.g.chitu@ewi.tudelft.nl
Web address: http://mmi.tudelft.nl/~alin

LEON J.M. ROTHKRANTZ received the MSc. degree in mathematics from the University of Utrecht, Utrecht, The Netherlands, in 1971, the Ph.D. degree in mathematics from the University of Amsterdam, Amsterdam, The Netherlands, in 1980, and the MSc. degree in psychology from the University of Leiden, Leiden, The Netherlands, in 1990. He is currently an Associate Professor with the Man-Machine Interaction Group, Mediamatics Department, Delft University of Technology, Delft, The Netherlands, since 1992. His current research focuses on a wide range of the related issues, including lip reading, speech recognition and synthesis, facial expression analysis and synthesis, multimodal information fusion, natural dialogue management, and human affective feedback recognition. The long-range goal of his research is the design and development of natural, context-aware, multimodal man–machine interfaces. Drs. Dr. Rothkrantz is a member of the Program Committee for EUROSIS.
Email: l.j.m.rothkrantz@ewi.tudelft.nl
Web address: http://mmi.tudelft.nl/~leon