AN AUDIO-VISUAL CORPUS FOR MULTIMODAL SPEECH RECOGNITION IN DUTCH LANGUAGE

Jacek C. Wojdeł, Pascal Wiggers, Leon J. M. Rothkrantz
Data and Knowledge Engeneering Group
Delft University of Technology
Mekelweg 4, 2628 BZ Delft, The Netherlands

ABSTRACT
This paper describes the gathering and availability of an audio-visual speech corpus for Dutch language. The corpus was prepared with the multi-modal speech recognition in mind and it is currently used in our research on lip-reading and bimodal speech recognition. It contains the prompts used also in the well-established POLYPHONE corpus and therefore captures the Dutch language characteristics with a reasonable accuracy.

1. INTRODUCTION

The availability of training and testing data is crucial when developing speech processing systems. There are already many commercially available speech corpora that contain audio data only. Certainly the TIMIT [1] database is one of the most often used in developing English language ASRs. For Dutch language, the POLYPHONE [2] dataset is comparably comprehensive. There is however lack of such datasets containing both audio and visual information. One of the few available resources is M2VTS database together with its successor XM2VTSDB [3]. This comprehensive audio-visual datasets were however designed and recorded with person identification applications in mind and therefore are not well suited for the development of multi-modal speech processing systems. Other available audio-visual datasets such as for example Tulips1 database [4] are too small and limited in the range of utterances to form a base for developing a bimodal speech recognizer. It was therefore crucial for our research in the area of lip-reading to gather our own dataset that would be appropriate for speech related research.

2. RECORDING REQUIREMENTS

From our earlier experiments with lip-reading we gathered some experience that allowed us to specify the requirements that the recorded data must satisfy in order to be suitable for developing an audio-visual ASR and/or lip-reading system. Those requirements in turn influenced both the content of the recordings and the physical setup during the sessions.

2.1. Audio requirements

In order to be useful for speech recognition the audio data must be of a reasonably high quality. We assume that the audio recordings sampled at 44kHz with 16-bit resolution would be sufficient for developing a speech recognizer. The speech recognition methods are sophisticated enough to allow for a limited noise in the recordings and the use of middle-class recording equipment. As the audio signal is substantially less storage expensive than the video signal, we decided to keep all of the audio data in uncompressed form, so that no signal degradation happens on the storage level.

2.2. Video requirements

Contrary to the audio data, it is not feasible to store the video in uncompressed form. We therefore decided to use the MPEG1 compression with a high bit-rate in order to make an optimal choice between image quality and file size. In order to speed up the development of the system we also decided that the camera will be focused only on the lower part of the face. This leads to the simplification of the liptracking process and allows to use lower video resolution. At a resolution comparable to the one used in M2VTS, we obtain a much finer detail in the mouth region images (see Fig. 1). Such a restricted field of view is of course not easily achievable in most of the real-life situations, it can be however justified in the development stage.
An additional concern when recording video data was the color reproduction of the used equipment. In the earlier experiments we have found out that the commercially available camcorders are very sensitive to the changes in the illumination conditions. It is inherent to all video coding standards (both analog and digital) to put more emphasis on image intensity than on its chromacity information. Therefore
we had to ensure that the recorded scene was well lit . and preferably with a natural light source.

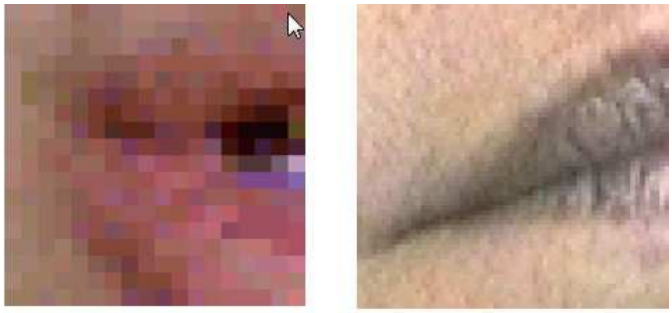

Fig. 1. A detail of mouth images in (a) M2VTS database and in (b) our recordings.

3. PROMPTS

The set of prompts that is used in our recordings is derived from the prompts recorded for the POLYPHONE [2] dataset. The POLYPHONE corpus consists of an extended set of telephone quality recordings of Dutch utterances. One of the major parts of this corpus consists of the natural language sentences that were were gathered from Dutch newspapers and grouped in sets of five in such a way that in each set each of the Dutch phonemes occurs at least once. We used those phonetically rich sets together with separate words, spelling examples and application specific utterances when preparing prompts for our recordings.

Our prompts collection is divided in 24 sections, each of them with the structure described later. Recording all of the 24 sections with each of the participants would not be really feasible as it would require almost a 2 hour sessions. All of the respondents agreed that one hour of recordings is already a hard experience. We constrained ourselves to the one hour sessions, which resulted in recordings between 10 and 14 sections of the prompt set. Because of the organizational issues such as introducing the respondent, resetting the setup etc. during a single hour of recordings we gathered between 25 and 45 minutes of actual material. Each section of the prompt set contains a fixed number of different utterances. The example section can be seen in Fig. 2. There are always 10 separate words, 10 phonetically rich sentences (2 sets from POLYPHONE), 3 ten-digit sequences, 4 spelled words and 5 application oriented utterances.

3.1. Words

The 10 words that open each section are meant for a single phoneme experiments. As it is hard for a non-trained subject to pronounce properly an isolated phoneme, we decided to choose the smallest possible words that contain each of

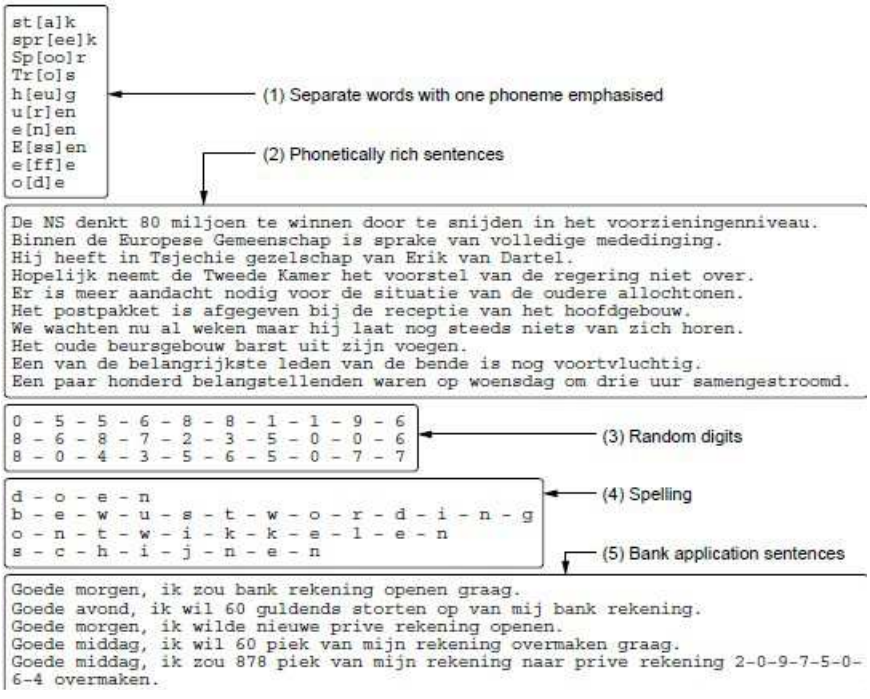

Fig. 2. Example set of prompts.

the phonemes. The letters corresponding to the selected phoneme are in each word highlighted on the respondent's display. The respondents were asked to pronounce this phoneme as well and clear as possible (possibly prolonging or stressing it). The words that have a vowel highlighted are in the form of CVC, CCVC or CVCC. In case of consonants we use words containing no more that two syllables and the consonant in question being in the middle of the word. The set of such words is pretty limited in Dutch language and therefore most of them occurred more than once in the whole prompt set.

3.2. Phonetically rich sentences

The 10 sentences in each section are formed of a pair of randomly chosen phonetically rich sets from POLYPHONE. Although the phonetically rich sentences guarantee the occurrence of each of the phonemes, they do not give the guarantee of providing the natural distribution of the phonemes. It could happen that the phoneme distributions in those sentences would be skewed in the direction of the least common phonemes. In order to check this we compared the phoneme histogram of the selected sentences to the histogram of the whole POLYPHONE data-set (see Fig. 3). The histogram of the whole POLYPHONE can be assumed to be a natural distribution of the phonemes in Dutch language as the forced utterances are only a small part of the whole data; most of the POLYPHONE contains spontaneous answers to a questionnaire.

3.3. Digits

This part of the prompt section is made up of 30 digits in total. They are presented to the respondent in 3 sequences,

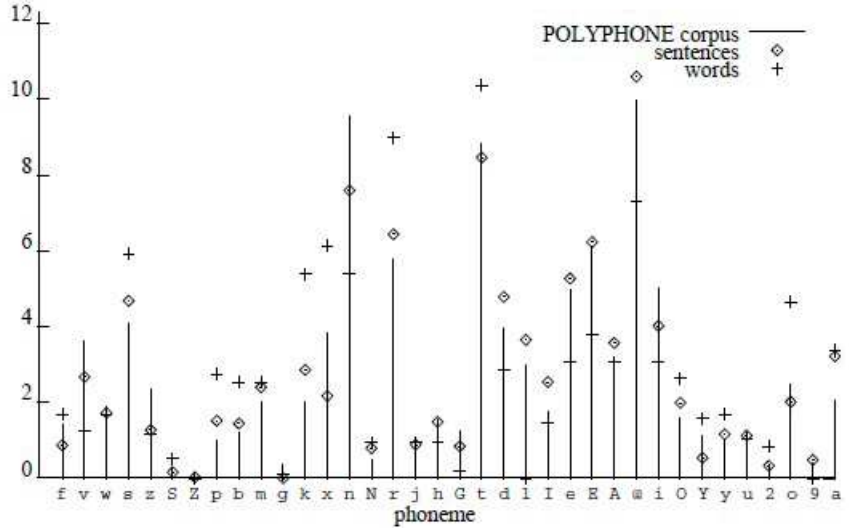

Fig. 3. Phoneme histogram for the POLYPHONE corpus,phonetically rich sentences in our prompt set and the phonemes selected in isolated words.

10 digits each. The digits were randomly generated and have uniform distribution in the whole prompt set. There was however no uniformity forced on a per-section basis. The digit recordings can be used in experiments with the limited vocabulary recognition.

3.4. Spelling

A spelling based recognizer can be deployed in case of e.g. a phone-book application, when transcribing all possible utterances would not be feasible [5]. Especially in this case spelling the name is a rather intuitive approach even in case of human-to-human communication. For this reason the spelling of 4 randomly chosen words is included in each section.

3.5. Application oriented utterances

In order to test for a real-world performance of the recognizer we need also some utterances with a constrained grammar and vocabulary. In this case we've chosen for a tele-banking application and prepared a simple grammar for the opening user utterance. The Hidden Markov Toolkit (HTK) was used to create a corresponding word-net and later to generate a set of random utterances from it. The grammar was prepared with recognition in mind, so some of the generated utterances are not grammatically correct. This however is not a big issue as we do not intend to deploy such a system, but rather just want to test the capabilities of bimodal speech recognition in a constrained grammar situation.

4. RECORDED DATA

At the current stage, we have recorded in total 87 sessions with 8 different respondents. This gives in total over 4

Table 1. Utterances recorded in the corpus.

	Normal	Slow	Whisp.	Total
Sections	58	22	7	87
Sentences	865	330	105	1300
Words	9380	3616	1153	14149
Words (sep.)	571	219	70	860
Digits	1683	627	218	2528

Table 1. Utterances recorded in the corpus.

hours of constant recordings. The recorded respondents were all native Dutch speakers. There are 7 male and only one female, the gender skew that couldn't be avoided in recordings of the volunteering students of Delft University of Technology. We asked the respondents to vary the speech rate during the recordings. Some of the recorded sections are marked as being "slowly spoken", which means that the respondents were asked to slow down the speech rate. A small amount of sessions were also recorded with respondents whispering the prompts in order to allow the investigation of this type of articulation as well. The total numbers of recorded utterances are summarized in Table 1.

5. DIGITAL STORAGE STRUCTURE

After recording all of the data on the mini-DV tapes, we have to store them in more convenient digital form. The storage structure is depicted in Fig. 4. Each of the recorded sessions was edited using a video editing software and cut into smaller sequences. The video sequences were then converted from a standard DV format to MPEG1 stream. Moreover, from all of the scenes the audio data was extracted and saved externally. Further, the proper transcriptions of the utterances were added.

The video sequences are in a half-PAL resolution (384x288) they are sampled at 25 frames per second and then saved with 600kbps bit-rate. This is a relatively high bitrate and together with a low amount of changes in the video it provides us with a fairly undistorted picture. The resulting files are stored on the CD-ROMs with the following directory structure:

Top-level directory – This directory contains only subdirectories with the names corresponding to the videotape on which the session was recorded.

Session directory – Each session directory contains two text files that describe the recording session and a number of section directories. The **info.txt** file is a strictly structured file with the basic information concerning the number of recorded sessions, speaker characteristics and other similar information. This file is mostly intended for automated use.

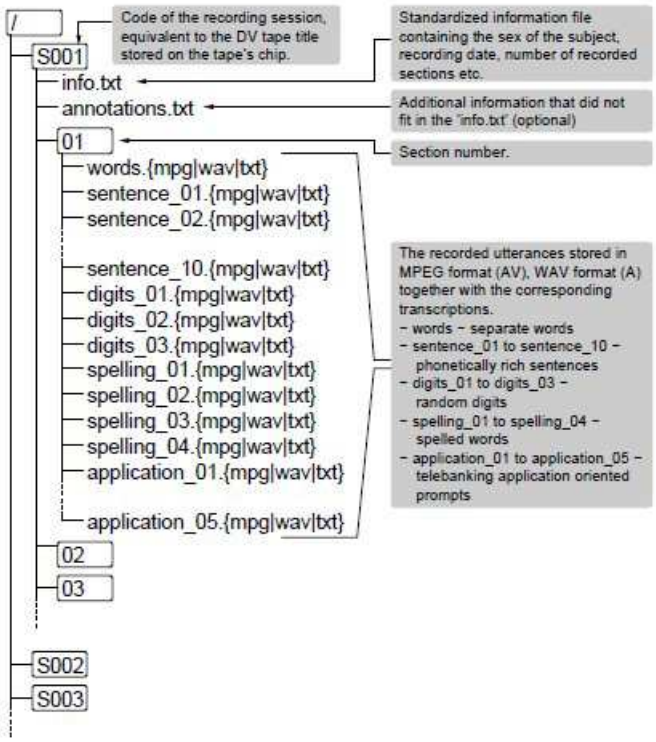

Fig. 4. The way in which the recorded data is stored on CD-ROMs.

The **annotations.txt** file on the other hand contains the verbal description of the recordings together with the description of any anomalies in the data and other things not captured in the **info.txt** file.

Section directory – The section directories are always numbered according to the section in the set of prompts that they contain. There are three types of files that are placed in here: MPEG1 encoded video sequences (file suffix: **mpg**), uncompressed audio in a standard wave format (**wav**) and the transcriptions of the utterances (**txt**). The different parts of the prompt set are stored in the files with the following names (without the type-dependent suffix):

words – the set of 10 phoneme specific words

sentence *number* – each of the 10 phonetically rich sentences

digits *number* – ten digits spoken with short pauses

spelling *number* – one of the 4 spelled words

application *number* – one of the 5 tele-banking related sentences

We have currently 7 CDs with 31 sections recorded in a group of 5 speakers (the single female speaker included). The available dataset is broad enough to allow us to development a person specific lip-reading system that recognizes strings of digits [6] and continuous audio-visual speech recognizer [7,8,9].

6. CONCLUSIONS AND FUTURE WORK

A substantial part of the presented audio-visual corpus is currently available for the research community. It is to our knowledge the first such corpus for Dutch language and one of the very few available for other languages. The amount of gathered data is sufficient for a development of the continuous speech recognition systems with medium vocabulary. The visual part of the recorded data has been prepared so that the amount of preprocessing needed to process it is minimal. There are no person- or head-tracking issues that need to be solved in advance. This assumption is not realistic in real-life settings, but allows us to separate the performance issues related to the other parts of the video-processing. The corpus must be fully segmented and stored on CDs in the future. This is a very time consuming task however and it will take additional time and resources. After putting all of the recorded data on CDs we will also consider extending it to a broader population of respondents. Another thing that can be considered is providing the labeling information for the data. The preliminary tests with automatic labeling with the ASR trained on a bigger audio dataset are very promising.

7. REFERENCES

[1] L. Lamel, R. Kassel, and S. Seneff, “Speech database development: Design and analysis of the acousticphonetic corpus,” in *Proc. DARPA Speech Recognition Workshop*, pp. 100–109, 1986.

[2] M. Damhuis, T. Boogaart, C. In ’t Veld, M. Versteijlen, W. Schelvis, L. Bos, and L. Boves, “Creation and analysis of the Dutch polyphone corpus,” in *Proceedings of the International Conference on Spoken Language Processing, ICSLP’94*, (Yokohama, Japan), pp. 1803– 1803, 1994.

[3] K. Messer, J. Matas, J. Kittler, J. Luettin, and G. Maitre, “XM2VTSDB: The extended M2VTS database,” in *Second International Conference on Audio and Videobased Biometric Person Authentication*, March 1999.

[4] J. R. Movellan, “Visual speech recognition with stochastic networks,” in *Advances in Neural Information Processing Systems* (T. L. G. Tesauro, D. Toruetzky, ed.), MIT Pess, Cambridge, 1995.

[5] R. van Vark, J. de Haan, and L. Rothkrantz, “A domain independent model to improve spelling in a web environment,” in *Proceedings of ICSLP 2000*, (Beijing, China), pp. 1081–1084, 2000.

[6] J. C. Wojdel and L. J. M. Rothkrantz, “Using aerial and geometric features in automatic lip-reading,” in *Proceedings of Eurospeech 2001 – Scandinavia*, pp. 2463– 2466, September 2001.

[7] P. Wiggers, J. C. Wojdeł, and L. J. M. Rothkrantz, “Medium vocabulary continuous audio-visual speech recognition,” in *Proceedings of ICSLP 2002*, 2002.

[8] P.Wiggers, J.C. Wojdel, and L.J.M.Rothkrantz. “Development of a speech recognize for the Dutch language” in Proceedings of the 7th annual scientific conference on webtechnology new media, communications and telematics theory, methods, tools and applications (EUROMEDIA), pp. 133-138, 2002.

[9] P.Wiggers, L.J.M.Rothkrantz. “Integration of speech recognition and automatic lip reading” in Text Speech and Dialogue, pp. 205-212. 2002