J Intell Inf Syst (2017) 49:167–192
DOI 10.1007/s10844-016-0438-z

An audio-visual corpus for multimodal automatic speech recognition

Andrzej Czyzewski[1] · Bozena Kostek[2] · Piotr Bratoszewski[1] · Jozef Kotus[1] · Marcin Szykulski[1]

Received: 5 July 2016 / Revised: 4 December 2016 / Accepted: 6 December 2016/
Published online: 7 January 2017

Abstract A review of available audio-visual speech corpora and a description of a new multimodal corpus of English speech recordings is provided. The new corpus containing 31 hours of recordings was created specifically to assist audio-visual speech recognition systems (AVSR) development. The database related to the corpus includes high-resolution, high-framerate stereoscopic video streams from RGB cameras, depth imaging stream utilizing Time-of-Flight camera accompanied by audio recorded using both: a microphone array and a microphone built in a mobile computer. For the purpose of applications related to AVSR systems training, every utterance was manually labeled, resulting in label files added to the corpus repository. Owing to the inclusion of recordings made in noisy conditions the elaborated corpus can also be used for testing robustness of speech recognition systems in the presence of acoustic background noise. The process of building the corpus, including the recording, labeling and post-processing phases is described in the paper. Results achieved with the developed audio-visual automatic speech recognition (ASR) engine trained and tested with the material contained in the corpus are presented and discussed together with comparative test results employing a state-of-the-art/commercial ASR engine. In order to demonstrate the practical use of the corpus it is made available for the public use.

Keywords MODALITY corpus · English language corpus · Speech recognition · AVSR

✉ Marcin Szykulski
marszyk@sound.eti.pg.gda.pl

[1] Faculty of Electronics, Telecommunications and Informatics, Multimedia Systems Department, Gdansk University of Technology, ul. Narutowicza 11/12, 80-233 Gdansk, Poland

[2] Faculty of Electronics, Telecommunications and Informatics, Audio Acoustics Laboratory, Gdansk University of Technology, ul. Narutowicza 11/12, 80-233 Gdansk, Poland

1 Introduction

Current advances in microelectronics make efficient processing of audio and video data in computerized mobile devices possible. Nowadays, most smartphones and tablet computers are equipped with audio-based speech recognition systems. However, when those functionalities are used in real environments, the speech signal can become corrupted, negatively influencing speech recognition accuracy (Trentin and Matassoni 2003). Besides co-occurring sound sources (background noise, other speakers), the performance can be degraded by reverberations or distortions in the transmission channel. Inspired by the human-like multimodal perception of speech described in the literature (e.g. by McGurk 1976), an additional information from the visual modality, usually extracted from a recording of speaker's lips, can be introduced in order to complement acoustic information and to mitigate the negative impact of audio corruption. Several researches have reported increased performance of multimodal systems when operating in noise compared to uni-modal acoustic speech recognition systems (Chibelushi et al. 1996), Kashiwagi et al. (2012), Potamianos et al. (2003), Stewart et al. (2014). Well established studies in the field of the Audio Visual Speech Recognition (AVSR) employ parametrization of facial features using Active Appearance Models (AAM) (Nguyen and Milgram 2009) and viseme recognition utilizing Hidden Markov Models (HMM) (Bear and Harvey 2016) or Dynamic Bayesian Networks (Jadczyk and Ziółko 2015). The most recent works employ Deep Neural Networks (DNN) (Almajai et al. 2016), Mroueh et al. (2015) and Convolutional Neural Networks (CNN) (Noda et al. 2015) serving as a front-end for audio and visual feature extraction. The usage of DNN or DNN-HMM (Noda et al. 2015), where the conventional Gaussian Mixture Model is replaced with DNN to represent connection between HMM states and input acoustic features, offers an improvement in terms of word accuracy over the baseline HMM. In the novel approach to visual speech recognition by Chung et al. (2016), Convolutional Neural Networks and a processing on the sentence level at both: learning and analysis phase rather than on the phoneme level were employed.

However, to design robust AVSR algorithms, a suitable speech material must be prepared. Because the process of creating a multi-modal dataset requires a considerable amount of time and resources (Chitu and Rothkrantz 2007), the number of available multi-modal corpora is relatively small compared to uni-modal corpora availability. Existing datasets often suffer from poor quality of video recordings included. It can be argued that for some cases, such as speech recognition employing low-quality webcams, the low-resolution multi-modal corpora better match the target applications. However, as video standards advance, their use is becoming more and more limited. Another problem of audio-visual speech corpora reported in research papers is that they are often not open to the public, or are commercial, thus researchers are forced to build their own datasets, especially in the case of national languages (Żelasko et al. 2016). Meanwhile, results achieved with some local datasets cannot be compared with results achieved with other ones, mostly because these corpora contain different material (also recorded in national language), a variety of audio-visual features and algorithms employed.

The multimodal database presented in this paper aims to address above mentioned problems. It is distributed free of charge to any interested researcher. It is focused on high recording quality, ease of use and versatility. All videos were recorded in 1080p HD format, with 100 frames per second. To extend the number of potential fields of use of the dataset, several additional modalities were introduced. Consequently, researchers intending to incorporate facial depth information in their experiments can do that owing to the second camera applied to form a stereo pair with the first one or by utilizing the recordings

from the Time-of-Flight camera. Investigating the influence of reverberation and noise on recognition results is also possible, because additional noise sources and a set of 8 microphones capturing sound at different distances from the speaker were used. Moreover, SNR (signal-to-noise ratio) values were calculated and made accessible for every uttered word (a detailed description of this functionality is to be found in Section 3.4).

The remainder of the paper is organized as follows: Section 2 provides a review of currently available audio-visual corpora. Our methods related to the corpus registration, including used language material, hardware setup and data processing steps are covered in Section 3, whereas Section 4 contains a description of the structure of the published database, together with the explanation of the procedure of gaining an access to it. Hitherto conceived use-cases of the database are also presented. Example speech recognition results achieved using our database, together with procedures and methods employed in experiments are discussed in Section 5. The paper concludes with some general remarks and observations in Section 6.

2 Review of audio-visual corpora

The available datasets suitable for AVSR research are relatively scarce, compared to the number of corpora containing audio material only. This results from the fact that the field of AVSR is still a developing relatively young research discipline. Another cause may be the multitude of requirements needed to be fulfilled in order to build a sizable audio-visual corpus, namely: a fully synchronized audio-visual stream, a large disk space, and a reliable method of data distribution (Durand et al. 2014).

As high-quality audio can be provided with relatively low costs, thus the main focus during the development of a AVSR corpus should be put on the visual data. Both: high resolution of video image and high framerate are needed in order to capture lip movement in space and time, accurately. The size of the speaker population depends on the declared purpose of the corpus - those focused on speech recognition, generally require employment of a smaller number of speakers than the ones intended for the use in speaker verification systems. The purpose of the corpus also affects the language material - continuous speech is favorable when testing speech recognition algorithms, while speaker verification can be done with separated words. Ideally, a corpus should contain both above types of speech. The following paragraphs discuss historic and modern audio-visual corpora in terms of: speaker population, language material, quality, and some other additional features. The described corpora contain English language material unless stated otherwise.

History of audio-visual datasets begins in 1984, when a first corpus was proposed by Petajan (1988) to support a lip reading digit recognizer. The first corpora were relatively low-scale, for example TULIPS1 (1995) contains short recordings of 12 speakers reading four first numerals in English (Movellan 1995). Bernstein Lipreading Corpus (1991) offers a more sizable language material (954 sentences, dictionary of 1000 words), however it contains recordings of only two speakers (Bernstein 1991).

One of the first more comprehensive data sets, namely DAVID-BT, was created in 1996 (Chibelushi et al. 2002). It is composed of 4 corpora with different research themes. The corpora focused on speech/speaker recognition consists of recordings of 123 speakers (31 clients with 5 recording sessions, 92 impostors with 1 recording session). The speech material of the database contains isolated numerals, the English-alphabet E-set, control commands for video-conferencing and 'VCVCV' (i.e. vowel-consonant-vowel-consonant-vowel, e.g. "awawa") nonsense utterances. The corpora are divided into subsets with

various recording conditions. The varying attributes include: visual background (simple or complex), lip highlighting, and profile shots.

The Multi Modal Verification for Teleservices and Security applications corpus (M2VTS) (Pigeon and Vandendorpe 1997), which was published in 1997, included additional recordings of head rotations in four directions - left to right, up and down (yaw, pitch), and an intentionally degraded recording material, but when compared to DAVID-BT, it is limited by small sample size and by the used language material, because it consists of recordings of 37 speakers uttering only numerals (from 0 to 9) recorded in five sessions.

M2VTS was extended by Messer et al. in 1999 (1999), and then renamed to XM2VTS. The sample size was increased to 295 subjects. The language material was extended to three utterances (including numerals and words) recorded in four sessions. The database was acquired under uniform recording conditions. The size of the database may be sufficient for identity verification purposes, but the still limited dictionary hinders potential research in the domain of speech recognition.

CUAVE (Clemson University Audio Visual Experiments), database designed by Patterson et al. (2002) was focused on availability of the database (as it was the first corpus fitting on only one DVD disc) and realistic recording conditions. It was designed to enhance research in audio-visual speech recognition immune to speaker movement and capable of distinguishing multiple speakers simultaneously. The database consists of two sections, containing individual speakers and speaker pairs. The first part contains recordings of 36 speakers, uttering isolated or connected numeral sequences while remaining stationary or moving (side-to-side, back-and-forth, head tilting). The second part of the database included 20 pairs of speakers for testing multispeaker solutions. The two speakers are always visible in the shot. Scenarios include speakers uttering numeral sequences one after another, and then simultaneously. The recording environment was controlled, including uniform illumination and green background. The major setback of this database is its limited dictionary.

The BANCA database (2003) (Bailly-Bailliére et al. 2003) was created in order to enable testing of multi-modal identity verification systems based on various recording devices (2 cameras and 2 microphones of varying quality were used) in different scenarios. Video and speech data were recorded for four European languages, with 52 speakers belonging to every language group (26 males and 26 females), in total of 208 subjects. Every speaker recorded 12 sessions, which contained 2 recordings each: one using speaker's true identity, and an informed imposter attack (the imposter knew the text uttered by the impersonated speaker). The sessions were divided into three different scenarios, controlled (high-quality camera, uniform background, low noise conditions), moderately degraded (cheap webcam, noisy office environment) and other adverse factors (high-quality camera, noisy environment). Uttered speech sequences are composed of numbers, speaker's name, address and date of birth. Inclusion of client-imposter scenarios among many different scenarios makes BANCA an useful database for developers of speaker verification systems.

The AVICAR ("audio-visual speech in a car") (Lee et al. 2004) database, published in 2004 by Lee et al., was designed with low-SNR audio-visual speech recognition in mind. Additional modalities were included in the setup in order to provide complementary information that could be used to mitigate the effects of background noise. The recording setup included a microphone array (containing 8 microphones) and a camera array composed of 4 cameras. The microphone array was used in order to allow the study of beamforming techniques, while the camera array enables the extraction of 2D and 3D visual features. The constructed recording setup was placed in a car. The recordings were made in different noise conditions - while the car was moving at 35 and 55 miles per hour and while idling. To

introduce increased levels of noise, the recordings in the moving car were repeated while the car windows were open. The released corpus contains recordings of 86 speakers (46 male, 40 female), including native and non-native English speakers. The language material uttered by every speaker in the corpus included isolated letters and numerals, phone numbers and sentences from the TIMIT (Garofolo et al. 1993) corpus. The diverse vocabulary allows for research in recognition of isolated commands and continuous speech. Biswas et al., successfully utilized the data from the AVICAR corpus in the audio-visual speech recognition system of their design, which was found to be more robust to noise than the one trained with audio features only (Biswas et al. 2015).

The aim of the database published by Fox et al. (2005), named VALID, was to highlight the importance of testing multi-modal algorithms in realistic conditions by comparing the results achieved using controlled audio-visual data with the results employing uncontrolled data. It was accomplished by basing the structure of the database on an existing database XM2VTS, and introducing uncontrolled illumination and acoustic noise to the recording environment. The database includes the recordings of 106 speakers in five scenarios (1 controlled, 4 real-world) uttering the XM2VTS language material. Visual speaker identification experiments carried out by the authors of the new database VALID highlighted the challenges posed by poor illumination., which was indicated by the drop of ID detection accuracy from 97.17 % (for controlled XM2VTS data) to 63.21 % (for uncontrolled VALID data).

Another attempt in expanding the XM2VTS corpus is DXM2VTS (meaning "damascened" XM2VTS), published in 2008 by Teferi et al. (2008). Similar to VALID, it attempts to address the limitations of XM2VTS stemming from invariable background and illumination. Instead of re-recording the original XM2VTS sequences in different real-life environments, the authors used image segmentation procedures to separate the background of the original videos, recorded in studio conditions, in order to replace it with an arbitrary complex background. Additional transformations can be made to simulate real noise, e.g. blur due to zooming or rotation. The database is offered as a set of video backgrounds (offices, outdoors, malls) together with XM2VTS speaker mask, which can be used to generate the DXM2VTS database.

GRID corpus (2006, Cooke et al. 2006) was designed for the purpose of speech intelligibility studies. Inclusion of video streams expands its potential applications to the field of AVSR. The structure of GRID is based on the Coordinate Response Measure corpus (CRM) (Bolia et al. 2000). Sentences uttered by the speakers resembling commands have the form of: "<command:4><color:4><preposition:4><letter:25><digit:10><adverb:4>" (e.g. "place blue at A 0 again") where the digit indicates the number of available choices. All 34 speakers (18 male, 16 female) produced a set of 1000 different sentences, resulting in the total corpus size of 34,000 utterances. The video streams were captured synchronously in an environment with uniform lighting and background. The authors presented an experiment in audio intelligibility employing human listeners, made with acquired audio recordings. However, the corpus can be used for ASR and AVSR research as well, owing to word alignments, compatible with the Hidden Markov Model Toolkit (HTK) (Young et al. 2006) format, supplied by the authors.

As a visual counterpart to the widely-known TIMIT speech corpus (Garofolo et al. 1993), Sanderson (2009) created the VIDTIMIT corpus in 2008. It is composed of audio and video recordings of 43 speakers (19 female and 24 male), reciting TIMIT speech material (10 sentences per person). The recordings of speech were supplemented by a silent head rotation sequence, where each speaker moved their head to the left and to the right. The rotation sequence can be used to extract the facial profile or 3D information. The corpus

was recorded during 3 sessions, with average time-gap of one week between sessions. This allowed for admitting changes in speakers' voice, make-up, clothing and mood, reflecting the variables that should be considered with regards to the development of AVSR or speaker verification systems. Additional variables are: the camera zoom factor and acoustic noise presence, caused by the office-like environment of the recording setup.

The Czech audio-visual database UWB-07-ICAVR (Impaired Condition Audio Visual speech Recognition) (2008) (Trojanová et al. 2008) is focused on extending existing databases by introducing variable illumination, similar to VALID. The database consists of recordings of 10000 continuous utterances (200 per speaker; 50 shared, 150 unique) taken from 50 speakers (25 male, 25 female). Speakers were recorded using two microphones and two cameras (one high-quality camera, one webcam). Six types of illumination were used during every recording. The UWB-07-ICAVR database is intended for audio-visual speech recognition research. To aid it, the authors supplemented the recorded video files with visual labels, specifying regions of interest (a bounding box around mouth and lip area), and they transcribed the pronunciation of sentences into text files.

IV2, the database presented by Petrovska et al. (2008), is focused on face recognition. It's a comprehensive multimodal database, including stereo frontal and profile camera images, iris images from an infrared camera, and 3D laser scanner face data, that can be used to model speakers' faces accurately. The speech data includes 15 French sentences taken from around 300 participating speakers. Many visual variations (head pose, illumination conditions, facial expressions) are included in the video recordings, but unfortunately, due to the focus on face recognition, they were recorded separately and they do not contain any speech utterances. The speech material was captured in optimal conditions only (frontal view, well-illuminated background, neutral facial expression).

The database WAPUSK20, created by Vorwerk et al. (2010), is more principally focused on audio-visual speech recognition applications. It is based on the GRID database, adopting the same format of uttered sentences. To create WAPUSK20, 20 speakers uttered 100 GRID-type sentences each of them recorded using four channels of audio and a dedicated stereoscopic camera. Incorporating 3D video data may help to increase the accuracy of lip-tracking and robustness of AVSR systems. The recordings were made under typical office room conditions.

Developed by Benezeth et al. (2011) the BL (Blue Lips) (Benezeth and Bachman 2011) database, as its name suggests, is intended for research in audio-visual speech recognition or lip-driven animation. It consists of 238 French sentences uttered by 17 speakers, wearing blue lipstick to ease the extraction of lip position in image sequences. The recordings were performed in two sessions, the first one was dedicated to 2D analysis, where the video data was captured by a single front-view camera. The second session, was dedicated to 3D analysis, where the video was recorded by 2 spatially aligned cameras and a depth camera. Audio was captured by 2 microphones during both sessions. To help with AVSR research, time-aligned phonetic transcriptions of the audio and video data were provided.

The corpus developed by Wong et al. (2011) UNMC-VIER (Wong et al. 2011), is described as a multi-purpose one, suitable for face or speech recognition. It attempts to address the shortcomings of preceding databases, and it introduces multiple simultaneous visual variations in video recordings. Those include: illumination, facial expression, head poses and image quality (an example combination: illumination + head pose, facial expression + low video quality). The audio part also has a changing component, namely the utterances are spoken in slow and in normal rate of speech to improve the learning of audio-visual recognition algorithms. Language material is based on the XM2VTS sentences (11 sentences used) and is accompanied by a sequence of numerals. The database includes

recordings of 123 speakers in many configurations (two recording sessions per speaker - in controlled and uncontrolled environment, 11 repetitions of language material per speaker).

The MOBIO database, developed by Marcel et al. (2012), is a unique audio-visual corpus, as it was captured almost exclusively using mobile devices. It is composed of over 61 h of recordings of 150 speakers. The language material included a set of responses to short questions, also responses in free speech, and pre-defined text. The very first MOBIO recording session was recorded using a laptop computer, while all the other data were captured by a mobile phone. As the recording device was held by the user, the microphone and camera were used in an uncontrolled manner. This resulted in a high variability of pose and illumination of the speaker together with variations in the quality of speech and acoustic conditions. The MOBIO database delivers a set of realistic recordings, but it is mostly applicable to mobile-based systems.

Audiovisual Polish speech corpus (AGH AV Corpus) (AGH University of Science and Technology 2014) is an interesting example of an AVSR database built for Polish language. It is hitherto the largest audiovisual corpus of Polish speech (Igras et al. 2012; Jadczyk and Ziółko 2015). The authors of this study evaluate the performance of a system built of acoustic and visual features and Dynamic Bayesian Network (DBN) models. The acoustic part of the AGH AV corpus is more thoroughly presented and evaluated in the paper by the team of the AGH University of Sciece and Technology (Żelasko et al. 2016). Besides the audiovisual corpus, presented in Table 1, authors developed various versions of acoustic corpora featuring the large number of unique speakers, which amounts to 166. This results in over 25 h of recordings, consisting of a variety of speech scenarios, including text reading, issuing commands, telephonic speech, phonetically balanced 4.5 h subcorpus recorded in an anechoic chamber, etc.

The properties of above discussed corpora, compared with those concerning our own corpus, named MODALITY, are presented in Table 1.

The discussed existing corpora differ in language material, recording conditions and intended purpose. Some are focused on face recognition (e.g. IV2) while others are more suitable for audio-visual speech recognition (e.g. WAPUSK20, BL, UNMC-VIER). The latter kind can be additionally sub-divided according to the type of utterances to be recognized. Some, especially early created databases, are suited for recognition of isolated words (e.g. TULIPS1, M2VTS), while others are focused on continuous speech recognition (e.g. XM2VTS, VIDTIMIT, BL).

The common element of all of the reviewed databases is the relatively low video quality. The maximum offered video resolution for English corpora is equal to 708×640 pixels. This resolution is still utilized in some devices (e.g. webcams), but as many modern smartphones offer the recording video resolution of 1920×1080 pixels, thus it can be considered as outdated. Another crucial component in visual speech recognition, the framerate, rarely exceeding 30 fps, reaching 50 fps in case of UWB-07-iCAV and AGH. Although some databases may be superior in terms of the number of speakers or variations introduced in the video stream (e.g. lighting), our audio-visual corpus (MODALITY) is to the authors' best knowledge, the first in case of English language to feature the full HD video resolution (1920×1080) with the superior 100 fps framerate. Additionally, for some speakers in the corpus, the Time-of-Flight camera was used, enabling the depth image for further analysis. The employed camera model is SoftKinetic DepthSense 325 which delivers the depth data at 60 frames per second and with spatial resolution of 320×240 pixels. Besides of depth recordings, the 3D data can be retrieved owing to stereo RGB cameras recordings available in the corpus.

Table 1 Comparison of existing databases (databases contain English language material unless stated otherwise)

Database	Year	# of spk.	Res.	Fps	Language material	Additional features
TULIPS1	1995	12	100 × 75	30 fps	numerals 1–4	no
DAVID	1996	123	640 × 480	30 fps	numerals, alphabet, nonsense utterances	varying background
M2VTS	1997	37	286 × 350	25 fps	isolated numerals 0–9	head rotations, glasses, hats
XM2VTS	1999	295	720 × 576	25 fps	3 sentences (numerals and words)	head rotations, glasses, hats
CUAVE	2002	30	720 × 480	29.97 fps	isolated or connected numerals (7000 utterances total)	simultaneous speech
BANCA	2003	52	720 × 576	25 fps	numerals, name, date of birth and address	controlled, degraded and adverse conditions, impostor recordings
AVICAR	2004	84	360 × 240	29.97 fps	Isolated numerals and letters, phone numbers, TIMIT sentences	automotive noise, microphone and camera array
VALID	2005	106	720 × 576	25 fps	same as XM2VTS	varying illumination and noise
GRID	2005	34	720 × 576	25 fps	1000 command-like sentences	no
DXM2VTS	2008	295	720 × 576	25 fps	same as XM2VTS	varying background, video distortions
VIDTIMIT	2008	43	512 × 384	25 fps	10 TIMIT sentences	office noise and zoom
UWB-07-iCAV	2008	50	720 × 576	max 50 fps	continuous Czech utterances	varying illumination and quality
IV2	2008	300	780 × 576 max	25 fps	15 French sentences	stereo frontal and profile views, iris images, 3D scanner data, head pose and illumination variations
WAPUSK20	2010	20	640 × 480	48 fps	100 GRID sentences	stereoscopic camera, office noise
BL	2011	17	640 × 480	30 fps	238 French sentences	depth camera, highlighted lips
UNMC-VIER	2011	123	708 × 640 max	29 fps	12 XM2VTS sentences	varying quality, speech tempo, expressions, illumination, head poses
MOBIO	2012	152	640 × 480	16-30 fps	32 questions	recorded on mobile devices, varying head pose and illumination
AGH AV Corpus	2014	20	1920 × 1080	25/50 fps	Isolated words, numerals	Polish language, audio: 16 bit/44.1 kHz, h.264 video codec
MODALITY	2015	35	1920 × 1080	100 fps	168 commands (isolated, sentences)	stereo camera, varying noise, microphone array, word SNR, additional depth camera

Those properties (especially the high framerate), are particularly important for the research of visual speech recognition. In available corpora, video streams with a frame rate of 25 fps are the most common. In such video streams, every video frame represents 40 ms of time. As shortest events in speech production can last a little over 10 ms (e.g. plosives) (Kuwabara 1996), such temporal resolution is insufficient to capture them. Our corpus provides a temporal resolution of 10 ms, which makes it well suited for the task of speech recognition based on lip features tracking. Owing to the inclusion of noisy recordings in our corpus, it is possible to examine whether the visual features improve the recognition rates in low-SNR conditions. Some selected speech recognition results achieved while using the corpus are presented in Section 5. The corpus can also be used to perform speaker verification using voice or face/lip features. Provided labels can be used to divide a speaker's recording into training and test utterance sets.

Additional innovative features of the MODALITY corpus include: supplying word-accurate SNR values to enable assessments of the influence of noise on recognition accuracy. The audio was recorded by a microphone array of 8 microphones in total, placed at three different distances to the speaker and, additionally, by a mobile device. A feature only rarely found in existing corpora, is that the whole database is supplied with HTK-compatible labels created manually for every utterance. Hence, the authors presume that these assets make the corpus useful for scientific community.

3 Corpus registration

3.1 Language material and participants

Our previous work on a multimodal corpus resulted in a database containing recordings of 5 speakers (Kunka et al. 2013). The recorded modalities included: stereovision and audio, together with thermovision and depth cameras. The language material contained in this database was defined in the studies of English language characteristics by Czyzewski et al. (2013), reflecting the frequentation of speech sounds in Standard Southern British. The resulting corpus could be used for research concerning vowel recognition.

The aim of the more recent work of the authors of this paper was to create an expanded corpus, with potential applications to audio-visual speech recognition field. The language material was tailored in order to simulate a voice control scenario, employing commands typical for mobile devices (laptops, smartphones), thus it includes 231 words (182 unique). The material consists of numbers, names of months and days and a set of verbs and nouns mostly related to controlling computer devices. In order to allow for assessing the recognition of both isolated commands and continuous speech, they were presented to speakers as a list containing a series of consecutive words, and sequences. The set of 42 sequences included every word in the language material. Approximately half of them formed proper command-like sentences (e.g. GO TO DOCUMENTS SELECT ALL PRINT), while the remainder was formed into random word sequences (e.g. STOP SHUT DOWN SLEEP RIGHT MARCH). Every speaker participated in 12 recording sessions. They were divided equally between isolated words and continuous speech. Half of the sessions were recorded in quiet (clean) conditions, but in order to enable studying the influence of intrusive signals on recognition scores, the remainder contained three kinds of noise (traffic, babble and factory noise) introduced acoustically through 4 loudspeakers placed in the recording room. To confirm the synchronization of modalities, every recording session included a hand-clap (visible and audible in all streams) occurring at the beginning and at the end of the session.

To enable a precise calculation of SNR for every sentence spoken by the speaker, reference noise-only recording sessions were performed before any speaker session. For synchronization purposes, every noise pattern was preceded by an anchor signal in a form of 1s long 1 kHz sine.

The corpus includes recordings of 35 speakers. The gender composition is 26 male and 9 female speakers. The corpus is divided between native and non-native English speakers. The group of participants includes 14 students and staff members of the Multimedia Systems Department of Gdańsk University of Technology, 5 students of the Institute of English and American Studies at University of Gdańsk, and 16 native English speakers. Nine native participants originated from the UK, 3 from Ireland and 4 from the U.S., whereas speakers' ages ranged from 14 to 60 (average age: 34 years). About half of the participants were 20-30 years old.

3.2 Hardware setup

The audio-visual material was collected in an acoustically adapted room. The video material was recorded using two Basler ace 2000-340kc cameras, placed at 30 cm from each other and 70 cm from the speaker. The speakers' images were recorded partially from the side at a small angle, due to the use of a stereo camera with the central optical axis directed towards the face center. The shift of the image depends on whether the left or right stereo camera image is used. The cameras were set to capture video streams at 100 frames per second, in 1080×1920 resolution. The Time-of-Flight (ToF) SoftKinetic DS325 camera for capturing depth images is placed at distance equal to 40 cm.

The audio material was collected from an array of 8 B&K measurement microphones placed in different distances from the speaker. First 4 microphones were located 50 cm from the speaker, next 2 pairs at 100 and 150 cm, respectively. An additional, low-quality audio source was a microphone located in a laptop placed in front of the speaker, at the lap level. The audio data was recorded using 16-bit samples at 44.1 kSa/s sampling rate with PCM encoding. The setup was completed by four loudspeakers placed in the corners of the room, serving as noise sources. The layout of the setup is shown in Fig. 1.

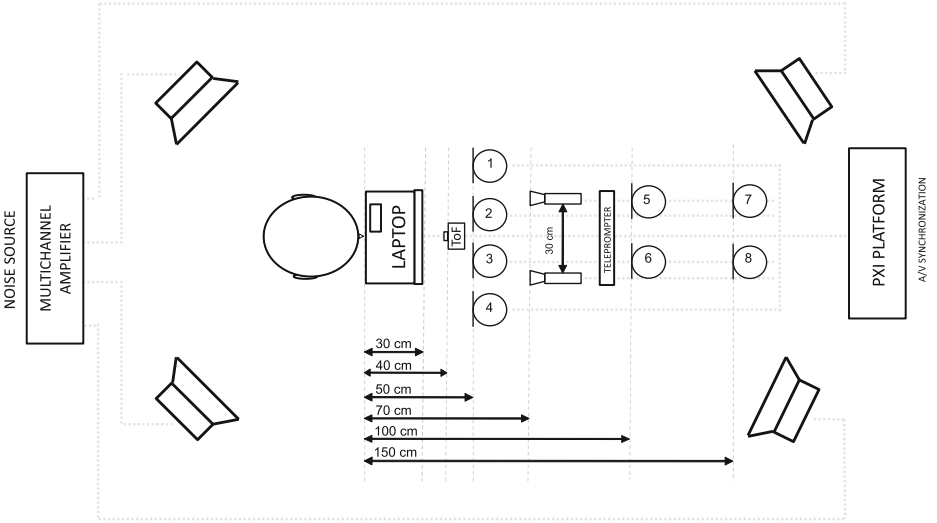

Fig. 1 Setup of the equipment used for recording of the corpus

To ensure a synchronous capture of audio and video streams, fast, robust disk drives were utilized, whereas the camera-microphone setup was connected to the National Instruments PXI platform supplied with necessary expansion cards and a 4 TB storage array. The registration process was controlled through a custom-built LabView-based application. The PC also ran a self-developed teleprompter application. The laptop computer and teleprompter did not obstruct the microphones and cameras in any way. The position of the loudspeakers, all microphones and cameras were permanently fixed during all recording sessions. The sound pressure level of the presented disturbing noise emission was also kept the same for all speakers.

3.3 Processing and labelling

As the raw video files consumed an extensive volume of space (about 13 GB of data for a minute-long recording of a single video stream) a need for a compression arose. Beforehand, an additional processing was needed in order to perform demosaicing of the original Bayer pattern images, which was performed using a self-developed tool for this purpose. The compression was done in ffmpeg using h.264 codec. The results were saved to '.mkv' container format with the size of almost 18 times smaller than the original files size. Some sample images are presented in Fig. 2. Additionally, the h.265 codec was used in order to reduce the amount of data needed to be downloaded by the corpus users. Therefore, the material is available in two versions: one encoded using h.264 and another one using h.265 codec. Authors decided to use two codecs as the second one is currently still less popular and its full implementation is still under development. However, as the h.265 codec is more future-oriented, thus the user is given a choice per coding type, entailing the file size. The depth data from Time-Of-Flight camera is recorded in RAW format and the sample images are presented in Fig. 3.

To facilitate the testing of audio-visual speech recognition algorithms, hand-made label files were created, to serve as ground truth data. This approach revealed also some additional advantages, especially that numerous minor mishaps have occurred during the recordings, including speakers misreading and repeating words or losing their composure (especially while reading random word sequences), instructions being passed to the speaker (e.g. please repeat) and pauses being made due to hardware functioning problems. The supplied label files include the position of every correctly-pronounced word from the set, formatted according to the HTK label format. This addition prevented from having to repeat the recording sessions after every mistake occurrence. Since the actual mistakes have not been

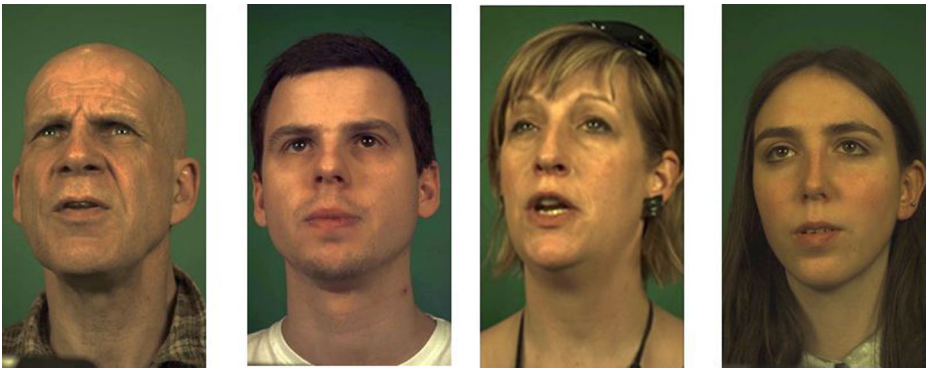

Fig. 2 Examples of video image frames from the MODALITY corpus

Fig. 3 Examples of depth image frames from the MODALITY corpus

removed from recorded material, it can be used to assess the effectiveness of disordered speech recognition algorithms (Czyzewski et al. 2003).

The file labeling was an extremely time-consuming process. The speech material was labeled at the word level. Initial preparations were made using the HSLab tool, supplied with HTK Speech Recognition Toolkit. However, after encountering numerous bugs and nuisances, it was decided to switch to a self-developed labeling application. Additional functionalities, such as easy label modification and autosave, facilitated the labeling process. Still, every hour of recording required about eleven hours of careful labeling work.

3.4 SNR calculation

The Signal-to-Noise ratio is the one of the main indicators used while assessing the effectiveness of algorithms for automatic speech recognition in noisy conditions. The SNR indicator is defined as the relation of signal power to noise power as expressed in the general form by (1):

$$SNR[dB] = 10\log_{10}\left(\frac{E_S}{E_N}\right) \tag{1}$$

where: E_S - energy of the speech signal, E_N - energy of the noise.

In order to accurately determine the SNR indicator according to the formula (1)), several steps were performed. First of all, during the preparation of the database, every type of disturbing noise was recorded separately. At the beginning of the noise pattern, a synchronizing signal (1 [kHz] sine of 1 [s] long) was added. The same synchronizing signal was played while making recordings of the speech signals in disturbed (noisy) conditions. Owing to this step, two kind of signals were obtained: disturbing noise only (E_N) and speech in noise (E_S+E_N). Both of those recordings include at the beginning the same synchronizing signal. After obtaining synchronization of the recordings, it was possible to calculate the energy of speech signal (E_S). A digital signal processing algorithm was designed for this purpose. The SNR calculations were performed in the frequency domain, for each FFT frame (index i in $E_{i,N}(f)$ and $E_{i,S+N}(f)$, denotes the i-th FFT frame of the considered signal). The applied algorithm can calculate instantaneous SNR ($SNR_i(i)$) based on formula (2):

$$SNR_i(i)[dB] = 10\log_{10}\left(\frac{E_{i,S}}{E_{i,N}}\right), \tag{2}$$

where: i - number of the FFT frame, $E_{i,S}$ - energy of the speech signal for i-th FFT frame, $E_{i,N}$ - energy of the noise for i-th FFT frame.

Based on energy components $E_{i,S}$ and $E_{i,N}$, the sum of energy of the speech signal $E_{w,S}$ and the sum of energy of the noise $E_{w,N}$ for a given word can be calculated using formulas (3) and (4):

$$E_{w,S}(j,k)[dB] = \sum_i^n E_{i,S}(j,k), \tag{3}$$

$$E_{w,N}(j,k)[dB] = \sum_i^n E_{i,N}(j,k), \tag{4}$$

where: j - number of the word spoken by k-th speaker, k - number of considered speaker, n - number of FFT frames for j-th word and k-th speaker (word boundaries were derived from the data contained in the label file - see next section for details).

Based on the sum of energy of noise and speech signal, the SNR for every recorded word (SNR_w) can be determined, according to formula (5):

$$SNR_w(j,k)[dB] = 10\log_{10}\left(\frac{E_{w,S}}{E_{w,N}}\right), \tag{5}$$

where: j - number of the word spoken by k-th speaker, k - number of considered speaker.
In the same way, it is also possible to calculate the average value of the SNR indicator for a given speaker (SNR_s), using formula (6):

$$SNR_s(k)[dB] = 10\log_{10}\left(\frac{E_{s,S}}{E_{s,N}}\right), \tag{6}$$

where: $E_{s,S}$ - the total energy of the speech signal for given speaker, $E_{s,N}$ - the total energy of the noise for given speaker.

Finally, it is possible to calculate the average SNR indicator (SNR_{AVG}) for all considered speakers and for given acoustic conditions using formula (7):

$$SNR_{AVG}[dB] = 10\log_{10}\left(\frac{1}{n}\sum_{k=1}^{n} 10^{\frac{SNR_S(k)}{10}}\right), \tag{7}$$

where: n - the number of considered speakers.

The block diagram illustrating the methodology of the SNR_i and SNR_w calculation is presented in Fig. 4. It shows the processing chain for a single microphone (analogous processing can be applied for all microphones in the array).

The proposed algorithm is based on simultaneous processing of two signals recorded during independent sessions. During the first session, only the acoustic noise was recorded. A recording of the speech signal disturbed by the noise was acquired during the second

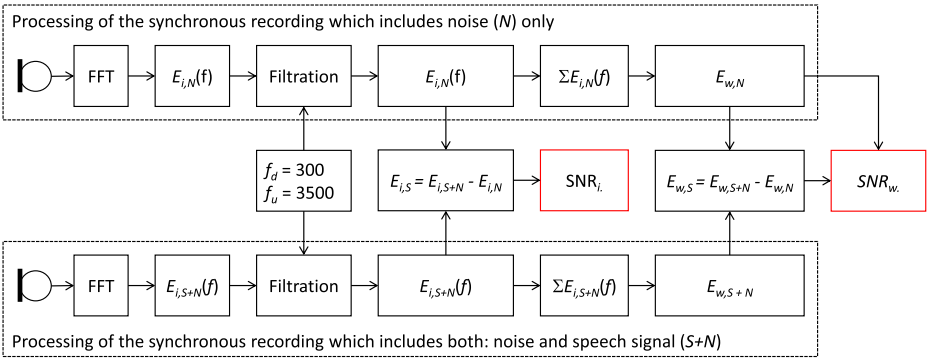

Fig. 4 Block diagram illustrating the SNR calculation methodology

session. After a manual synchronization of the signals, the energy of the signal E_N and noise E_S in the frequency domain can be calculated. The window length for the instantaneous SNR calculation was the same as the FFT frame and was equal to 4096 samples. The sampling rate for the acoustical signals was equal to 44100 Sa/s. Moreover, the calculation of the SNR value can be performed for the determined frequency range. In our corpus we provide two versions of SNR data. The first one represents the results of SNR calculation limited to the frequency range from 300 Hz (f_l - lower frequency limit) up to 3500 Hz (f_u - upper frequency limit) which corresponds to the traditional telephone bandwidth, whereas the second version was calculated for the full frequency range of human hearing (20 Hz - 20 kHz). Both versions are available in MODALITY downloadable assets. Based on the timestamps contained in the label file, it is possible to determine the SNR value for every spoken sentence according to formula (5) and average SNR value for considered speaker according on the basis of formula (6). These calculations were performed for all speakers and for all microphones in the array. In Fig. 5 the graphical presentation of the SNR_i and SNR_w. calculation results for selected speakers were depicted. Moreover, the energy of the speech and noise expressed in dB were also shown.

Based on the acoustic energy of speech (expressed in [dB]) and SNR_w calculated for every spoken word, the distribution of levels of the given indicator was calculated (for all speakers) as is presented in Fig. 6. We can observe that the disturbing noise causes a shift of the SNR histogram by 18.8 dB towards lower values. Moreover, due to the Lombard effect occurrence, the disturbing noise induces change in the speakers' voices, resulting mainly in louder speech utterances (Lane and Tranel 1971, 1993; Vlaj and Kacic 2011).

The average SNR value for clean conditions in the frequency range from 300 Hz up to 3500 Hz was equal to 36.0 dB. For noisy conditions the average SNR was equal to 17.2 dB. Calculation results of the average speech level for clean conditions and for noisy conditions were respectively: 66.0 dB and 71.7 dB. It means that during the recording in noisy conditions acoustic energy emitted by the speakers was 3.7 times greater than during clean conditions.

Information on SNR values described in this section (calculated for every audio file in the corpus) are included in a text files supplementing the corpus and are available for download from the MODALITY corpus homepage.

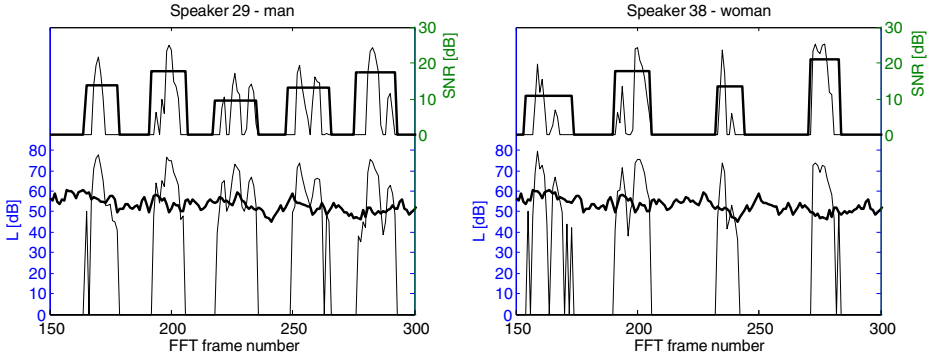

Fig. 5 Graphical presentation of the SNR_i and SNR_w. calculation results for selected speakers. *Bottom curves* present $E_{i,N}$ (*bold line*) and $E_{i,S}$ both expressed in [dB]. Upper curves present SNR_i (SNR for *i*-th FFT frame) and SNR_w (SNR value calculated for spoken word). Speech recordings were made in noisy conditions

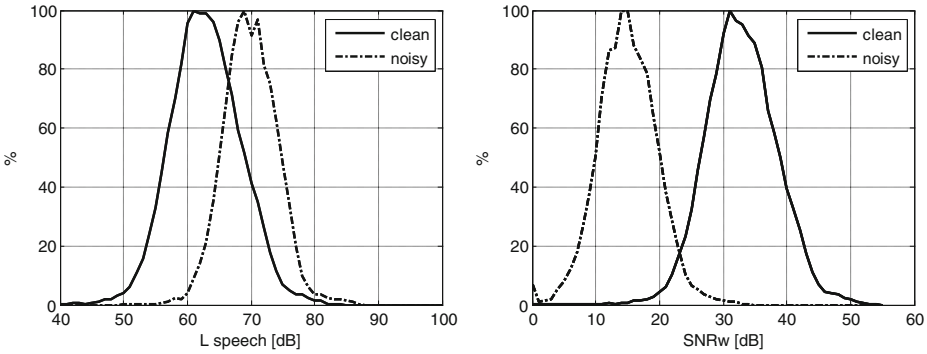

Fig. 6 Histogram of the speech level (*left*) and histogram of SNR values (*right*) calculated on the basis of SNRw, determined for every spoken words by all considered speakers stressed out in Section 5.2. Results obtained for mic. 2 (see Fig. 1), babble noise

3.5 Naming convention and utilized media formats

For every speaker 144 files were generated (9 audio files, 2 video files, 1 label file per 12 recording sessions), which were named according to the following principle:

SPEAKERNO_SESSIONNO_MODALITY.FORMAT

The file naming convention is presented in Table 2. For example:

SPEAKER24_S5_STRL.mkv

is a file containing the fifth session of sequence recording in noisy conditions of the speaker No. 24 by the left stereo camera.

The audio files use the Waveform Audio File Format (.wav), containing a single PCM audio stream sampled at 44.1 kSa/s with 16-bit resolution. The video files utilize the Matroska Multimedia Container Format (.mkv) in which a video stream in 1080p resolution, captured at 100 fps was used after being compressed employing both the h.264 and h.265 codecs (using High 4:4:4 profile). The .lab files are text files containing the information on word positions in audio files, following the HTK label format. Each line of the .lab file contains the actual label preceded by start and end time indices (in 100 ns units) e.g.:

12396200001244790000 FIVE

Table 2 Naming rules of the corpus files

No.	Speakers	Sessions		
	1–42	1–3 (quiet conditions) 4–6 (noise)		
Session	Separated commands	Command sequences		
	C	S		
MODALITY	Microphone array	Laptop microphone	Left camera	Right camera
	AUD1-8	LAPT	STRL	STRR
Format	Audio files	Video files	Label files	
	.wav	.mkv	.lab	

which denotes the word “five”, occurring between the 123.962 s and 124.479 s of audio material.

4 Access to corpus

The corpus has been made publicly available through the servers of the Gdansk University of Technology. The database is accessible at the web address: http://www.modality-corpus.org. The access is free, but the user is obliged to accept the license agreement. The web service overview page is presented in Fig. 7. The website is divided into four subpages:

- Home
- License
- About
- Explore corpus

Home subpage is an introductory page containing a short summary of the offered corpus. License explains the conditions under which the usage of the corpus is allowed. Additional information concerning the corpus can be found on the About subpage. The list of available files is located on the Explore corpus subpage. The access to the file list is granted only after accepting the license agreement. The subpage provides users with information on every speaker contained in the corpus, including gender, age & photo linked to a list of files corresponding to speaker’s recordings.

The material collected in the corpus uses a considerable amount of disk space (2.1 TB for h.264 codec, 350GB for h.265 codec). To give users the freedom to choose only the

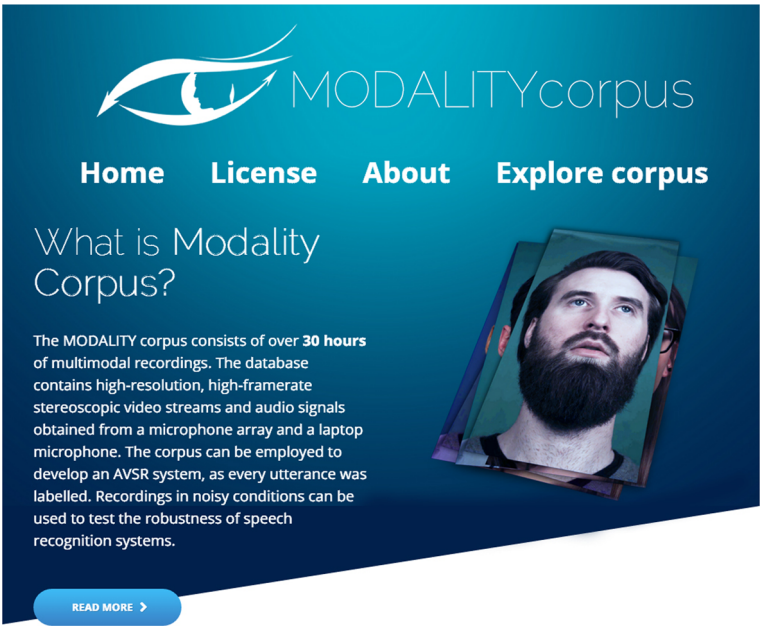

Fig. 7 Homepage of modality-corpus.org

recordings they need, the files of every recording session were placed in separate .zip files. The corpus was structured according to the speakers' language skills. Group A (16 speakers) consists of recordings of native-speakers. Recordings of non-natives (Polish nationals) were placed in Group B. The group of non-natives included 5 English language students and 14 faculty students and staff members.

5 Experimental results of speech recognition

In order to validate data, gathered in the developed corpus, the experiments in speech recognition were performed. A comparison of a commercially available state-of-the-art ASR engine with a self-developed ASR engine preceded planned experiments. The self-developed ASR was implemented utilizing HTK toolkit based on Hidden Markov Models (HMM). It makes possible adding visual features besides acoustic ones. The Mel-Frequency Cepstral Coefficients (MFCC) were employed in the acoustic speech recognition mode. They were complemented by vision-based parameters calculated owing to self-developed parametrization methods of the visemes in the AVSR mode. The conducted experiments consisted of solely audio-based or combined audio-visual speech recognition attempts as described in the following subsections.

5.1 Employed features and methods

In our research multiple feature extraction methods were utilized. In the acoustic layer of the AVSR system the standard MFCC features were used. The features were extracted from consecutive 10 ms – long fragments of the input signal. The Hamming window was applied to the analysis frame corresponding to the speech duration of 25 ms. The preemphasis with the coefficient equal to 0.97 was used at the acoustical signal preprocessing stage. For MFCC calculation, 13 triangular bandpass filters were used. The coefficients are calculated using the formula known in the literature (Young et al. 2006) which is directly derived from the work of Davis and Mermelstein (1980) as is presented in (8):

$$C_i = \sqrt{\frac{2}{N}\sum_{k=i}^{N} X_k \cos\left[\frac{\pi i}{N}(k-0.5)\right]}, i = 1, 2\cdots M, \tag{8}$$

where: N is the number of subchannels, X_k, $k = 1, 2, \cdots, M$ represents the log-energy output of the k-th filter.

Considering M = 13 subbands of the spectrum together with delta and delta-delta features results in the total number of 39 acoustic parameters used for this work. Multiple visual features are provided within the MODALITY corpus. All utilized visual features are based on characteristics of each speaker's lips region which had to be detected in advance to parametrization. Lip detection is performed using Active Appearance Models (AAM) algorithm. Nguyen and Milgram (2009). AAM algorithm is a general utility for statistical parametrization of objects based on Principal Component Analysis (PCA). For the detection step, the individual AAM model for each speaker was prepared consisting of 25 points on the speaker's face, including: 11 points denoting outer lip contour, 6 for an inner lip contour and 4 additional points for each nostril, which significantly improved the stability of detected lip shapes, especially when speakers' mouth was closed (Dalka et al. 2014). Individual AAMs used for lips detection were trained on 16 frames of gathered video material using manual annotation of mentioned 25 points for each speaker. It was the sufficient

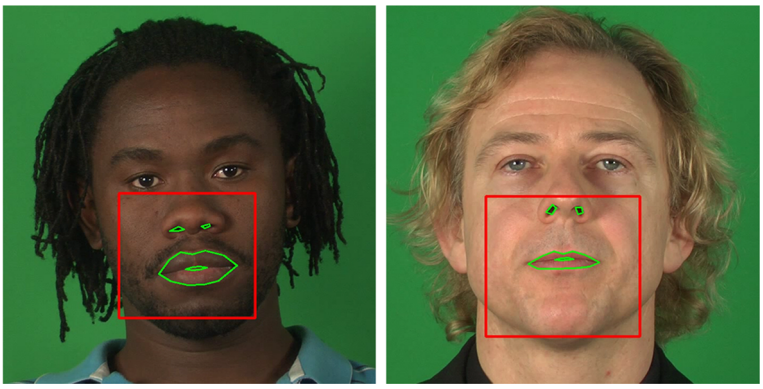

Fig. 8 Lip detection results obtained using AAM -based algorithm

amount of training frames for models to enable automatic and accurate lip detection on formerly unseen video images. The result of lips detection process is visible in Fig. 8. The AAM besides of lip detection is also used for the purpose of parametrization. An additional, general for all speakers, Active Appearance Model was created which uses automatic lip detection as a starting point. During the AAM learning, all lip shapes are aligned to one, normalized sample shape and then the mean shape is acquired. Textures are warped to this mean shape and then normalized in order to become invariant to lighting conditions. Subsequently, the PCA is performed for shapes and textures, independently. Results of PCA are truncated based on eigenvectors cumulative variance. Lip shape x can be approximated as the sum of mean shape $\bar{x}$ and the linear combination of the eigenvectors of shapes ϕ_s with the highest variation as in (9):

$$x = \bar{x} + \phi_s \cdot b_s, \tag{9}$$

where: b_s is corresponding to vector of coefficients (AAM-Shape parameters).

For texture parameters the feature extraction process is similar, namely the lip region texture g may be approximated as the sum of mean texture $\bar{g}$ and the linear combination of the eigenvectors of the texture ϕ_g revealing the highest variation (10):

$$g = \bar{g} + \phi_g \cdot b_g, \tag{10}$$

where b_g denotes a vector of coefficients (AAM-Texture parameters).

The shape parameters indicate information concerning lip arrangement, whereas texture parameters determine, for instance, tongue or teeth visibility. In the MODALITY corpus the AAM-Combined parameters are provided, which regard both the shape and texture information of the modeled object. Those parameters are obtained by concatenating AAM-Texture and AAM-Shape parameters and then performing PCA in order to remove correlation between both representations. The detailed description of the AAM algorithm implementation can be found in related publications (Dalka et al. 2014; Kunka et al. 2013).

Further visual parameters provided in the MODALITY corpus are named Dct-Inner and Dct-Outer. Dct-Inner parameters denote 64 DCT coefficients calculated from the bounding rectangle placed on the inner lips contour which is linearly interpolated to the region of 32x32 pixels. The DCT transform is computed from the luminance channel L of LUV color space as in (11):

$$DCT = X_N \cdot L \cdot (X_N)^T, \tag{11}$$

where:

$$X_N(j,k) = \sqrt{\frac{\alpha_j}{N}} \cos\left(\frac{\pi(2k+1)j}{2N}\right), \alpha_j = \begin{cases} 1, & \text{if } j = 0. \\ 0, & \text{if } j > 0. \end{cases}, j,k = 0, \cdots, N-1, \quad (12)$$

j and k are the actual analyzed pixel coordinates, and N is equal to 32.

Similarly, the DCT-Outer is calculated, besides the outer lips contour instead of the inner contour is enclosed by the bounding rectangle.

Furthermore, the luminance histogram parameters are provided. Both Histogram-Inner and Histogram-Outer parameters represent the 32-element luminance histogram in which the bins are evenly distributed over the luminance variation range. Histogram-Inner denotes the analysis bounding rectangle placed on the inner lips region, whereas Histogram-Outer represents the outer lips region. Moreover, the vertical lips profile is extracted in following manner: the rectangular bounding box encloses the outer lip region or the inner lip region, then it is scaled using linear interpolation to 16-pixel height and the for each row of the rectangle the mean value of R channel from RGB color space is calculated, resulting in VerticalProfile-Outer or Vertical-ProfileInner parameters.

Finally, statistics of Co-Occurrence Matrix (GCM-Inner and GCM-Outer) of lips pixels in 4 directions are used. The Co-Occurrence Matrix C is defined as in (13):

$$C(i,j) = \sum_{p=1}^{n} \sum_{q=1}^{m} \begin{cases} 1, & \text{if } I(p,q) = i \text{ and } I(p+\Delta x, q+\Delta y) = j. \\ 0, & \text{otherwise} \end{cases} \quad (13)$$

where: i and j are the image intensity values, (p,q) are the coordinates, n and m define the size of the image I and $(\Delta x, \Delta y)$ is the offset. For feature extraction, the region of interest is placed either on the outer lip contour (GCM-Outer) or on the inner lip contour (GCM-Inner). The C matrix is computed in four θ directions: 0, 45, 90 and 135 degrees. Those matrices are calculated for L, U components of the LUV color space and for vertical derivative of luminance L' of the image. Color depth is quantized to 16 levels. Hence, the resulting Co-Occurrence Matrix C is of size 16×16. The C matrix is then normalized and symmetrized resulting in a new matrix denoted as P. The following statistical descriptors of the matrix P are used as the visual parameters and then are calculated employing formulas (14–18):

$$K = \sum_{i,j} P_{i,j}(i-j)^2, \quad (14)$$

$$E = \sqrt{\sum_{i,j} P_{i,j}^2}, \quad (15)$$

$$\mu = \mu_i = \mu_j = \sum_{i,j} i \cdot P_{i,j} = \sum_{i,j} j \cdot P_{i,j}, \quad (16)$$

$$\sigma = \sigma_i = \sigma_j = \sqrt{\sum_{i,j} (i-\mu_i)^2 \cdot P_{i,j}} = \sqrt{\sum_{i,j} (j-\mu_j)^2 \cdot P_{i,j}}, \quad (17)$$

$$Corr = \sum_{i,j} P_{i,j} \frac{(i-\mu_i)(j-\mu_j)}{\sqrt{\sigma_i \sigma_j}}, \quad (18)$$

where: K, E, μ σ $Corr$ denote respectively: contrast, energy, mean value, standard deviation, correlation, and (i, j) are the coordinates in the matrix $P_{i,j}$. The resulting vector of parameters is of the size 60 (3 images $(L, U, L') \times 4$ matrices $(\theta = 0, 45, 90, 135) \times 5$ statistical descriptors).

In Table 3 the list of visual features described above is shown accompanied with each vector size. All described parameters are provided with the MODALITY corpus as .csv format files, thus they can be used for further research.

Table 3 List of available visual parameters in MODALITY corpus

Visual parameter	Vector size
AAM-Combined	40
AAM-Shape	22
AAM-Texture	58
Dct-Inner	64
Dct-Outer	64
GCM-Inner	60
GCM-Outer	60
Histogram-Inner	32
Histogram-Outer	32
VerticalProfile-Inner	16
VerticalProfile-Outer	16

5.2 ASR experimental methodology

Triphone-based left-right Hidden Markov Models (HMM) with 5 hidden states were used in the process of speech recognition. The model training was performed with the use of the HTK toolkit. The unigram language model was employed, meaning that every word symbol is equally likely to occur. In Fig. 9 the general lay-out of the speech recognition setup is presented. When both audio and visual features are used, they are concatenated into one fused vector of features (i.e. early integration) and then used for the model training and speech decoding tasks. The same HMM structure is used for audio and audio-visual speech recognition, however, for audio ASR the 39 MFCC features were provided to train the HMM models, whereas in case of audio-visual ASR the 39 MFCC and 10 AAM-Shape features were used. The 22 parameter AAM-Shape vector was truncated to the first 10 parameters. The AAM parameters in the provided vector are sorted from highest to lowest variance.

The word recognition process is based on the Bayesian theory, thus it requires the calculation of the maximum *a posteriori* probability (MAP) derived from the work of Young et al. (2006) and adapted to the problem of audio-visual speech recognition as in (19):

$$W = \arg\max_i P(W_i|O^{av}) = \frac{P(O^{av}|P(W_i)}{P(O^{av})}, \tag{19}$$

where: W represents the recognized word, W_i is the i-th word in a training data, O^{av} represents the sequence (vector) of combined both acoustic and visual features that can be

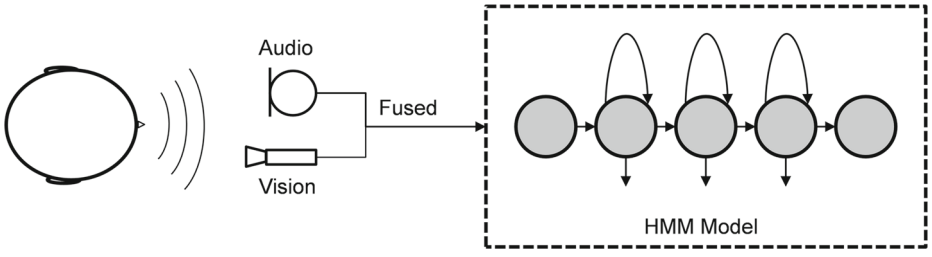

Fig. 9 Bi-modal speech recognition model

replaced in turn with O^a (acoustic sequence of observations) or O^v (visual sequence of observations), based on the actual model taken into consideration.

For the evaluation purposes the Word Error Rate (WER) metric is introduced (Park et al. 2008). According to WER definition, the recognition results are evaluated in terms of deletion, substitution and insertion errors, hence WER is calculated as in formula (20):

$$WER = \frac{D + S + I}{H + D + S} \cdot 100\ \% \tag{20}$$

where: H is the number of correctly recognized words, D is the number of deletions, S is the number of substitutions and I is the number of insertions.

The evaluation of implemented AVSR system was performed on recordings contained in the Modality corpus in leave-one-out (speaker independent) cross validation manner as shown in Table 4 bringing results described in the following subsection of the article. Values in Table 4 follow the terminology used in the Modality corpus as presented in Table 2 with more details. The test speaker was excluded from HMM models trainings and then the procedure was repeated until all speakers were tested. In case of testing in the presence of noise the C4 audio recordings were used.

5.3 Recognition results

The authors performed numerous experiments employing the multimodal corpus presented in this paper. The main goal of the research was to examine the role of the additional visual modality in the automatic speech recognition in difficult acoustic conditions, where the recognized speech was accompanied by babble, factory or street noise. The best performance of the above outlined AVSR system in noisy environment was achieved using AAM-Shape features of the length equal to 10 coefficients combined with 39 MFCC acoustic features. The mean WER results for clean speech and for speech distorted by babble noise, tested for all speakers represented in the corpus are presented in Table 5. Results indicate that babble noise dramatically worsens the accuracy of speech recognition (WER increases from 21 % to 51 %) and by addition of visual features to feature vector the accuracy of the proposed system increases by 5 percentage points (WER decreases by 5 %).

In Table 6 a comparison between speech recognition accuracy of the self-developed AVSR system and of the state-of-the-art, commercial ASR system incorporated into the Intel RealSense technology is presented. Both speech recognition systems were tested on the MODALITY corpus. The experimental speech material consisted of isolated commands separated by a short pause of 400ms. The RealSense engine has vast possibilities of customization - it can operate in an unconstrained mode with built-in large dictionary and

Table 4 ASR Experiments configuration

Configuration parameter	Used data set
Train speakers	1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 21 22 23 25 26 27 28 29 30 31 32 33 34 35 36 37 38 39 40 41
Test speaker	42
Training sessions	C1, C2, C3
Test sessions	C1, C2, C3, (C4)
Microphone	AUD2

Table 5 Accuracy of developed AVSR systemn

Noise	Acoustic features	Visual features	WER[%]
none	MFCC(39)	none	21
babble	MFCC(39)	none	51
babble	MFCC(39)	AAM-Shape(10)	46

language model (denoted as RS-Unconstr. In Table 6), as well as with limited word list and a self-prepared grammar. In the performed experiment, denoted as RS-Constr., the word list was limited to 184 speech commands recorded in the MODALITY corpus. Furthermore, the acoustic model is switchable to serve different languages. In the experiment two languages were tested, namely: British English (EN-GB) and American English (EN-US).

The results of speech recognition accuracy on the MODALITY speech corpus indicate the analogous behavior of both: the state-of-the-art ASR and the self-developed ASR system when noisy speech is introduced. Speech accompanied with babble noise is one of the most difficult cases for ASR systems to operate with, as it causes the occurrence of insertion and substitution errors in the recognized speech, leading to an increase of WER which is visible from achieved results. Limiting the list of possible word outputs (RS-Constr.) results in a significant decrease of WER when the isolated commands are recognized. Furthermore, the WER results for EN-GB are slightly better (i. e. lower) what can be explained by the fact of having native British English speakers in the MODALITY corpus. No WER results for audio video speech recognition in case of RealSense engine can be provided for comparison with our method as it is a closed solution accepting only audio stream as an input.

The remaining experiments and studies are thoroughly described in related publications. The influence of desynchronization between audio and visual features streams and video framerate was examined and described in the publication of Lopatka et al. (2014). Authors showed that the ± 100 ms offset between streams is acceptable in terms of early integration AVSR systems. Bratoszewski et. al. provided the study in the field of distant speech recognition (Bratoszewski et al. 2015). The experiments conducted in the cited work show the significance of proper HMM modeling based on the environment in which the ASR system is dedicated to operate. Authors proposed also a solution in which the ASR engine would dynamically switch the utilized acoustic models based on characteristics of the acoustic environment. Lopatka et al. examined the method of spatial filtration of speech signal in ASR system in noisy acoustic conditions, utilizing herein described speech corpus (Lopatka et al. 2015).

Table 6 ASR systems WER scores comparison on MODALITY speech corpus

Conditions	MODALITY	RS-Unconstr.	RS-Constr.	RS-Unconstr.	RS-Constr.
Clean	21	41.8	21.9	33.9	17.8
Noisy	51	61.5	49.2	54.5	41.3
Avsr	46	–	–	–	–
Dictionary	–	EN-US	EN-GB		

6 Summary and conclusions

A new multimodal English speech corpus was developed and presented. Owing to its open access it can be used for researching and developing audio-visual speech recognition systems. The MODALITY corpus enables researchers to test their AVSR algorithms using the video material recorded with a better quality then it was hitherto offered by available corpora. It also offers an opportunity to test the robustness of recognition algorithms, thanks to the inclusion of recordings made in challenging conditions (presence of acoustic noise, non-native speakers employment). The authors expect that the AVSR research progress will benefit from sharing the corpus among the research community free of charge. It is visible from preliminary experiments that in the environment where speech is accompanied by acoustic noise, the addition of visual parameters results in an improvement of automatic speech recognition accuracy and in lowering of WER. Further data analysis of provided audio and visual parameters in the MODALITY corpus may lead to creating of AVSR systems that will be more robust to noise than those representing state-of-the-art, commercially available ASR engines, based on the acoustic modality, only.

The future work related to a corpus extension is planned. The MODALITY corpus would benefit from extending it with additional recordings made with the same setup, containing some of the language material present in older, but popular corpora (e.g. VIDTIMIT, GRID). Hence, results achieved with the new, high-quality corpus might be compared with previous AVSR research achievements. The corpus could also be improved by adding some variations to the video modality (e.g. head poses, changing lighting conditions). Moreover, the comparison between results of the proposed classification methodology and several state-of-the-art classifiers is envisioned.

Acknowledgments Research partially sponsored by the Polish National Science Centre, Dec. No. 2015/17/B/ST6/01874.

The authors would like to thank Mr. Pawel Spaleniak for his help in developing the MODALITY corpus website.

References

AGH University of Science and Technology (2014). Audiovisual Polish speech corpus. http://www.dsp.agh.edu.pl/en:resources:korpusav, accessed: 2016-11-29.

Almajai, I., Cox, S., Harvey, R., & Lan, Y. (2016). Improved speaker independent lip reading using speaker adaptive training and deep neural networks. In *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 2722–2726). doi:10.1109/ICASSP.2016.7472172.

Bailly-Bailliére, E., Bengio, S., Bimbot, F., Hamouz, M., Kittler, J., Mariéthoz, J., Matas, J., Messer, K., Popovici, V., Porée, F., Ruiz, B., & Thiran, J.P. (2003). The BANCA Database and Evaluation Protocol. doi:10.1007/3-540-44887-X_74.

Bear, H.L., & Harvey, R. (2016). Decoding visemes: Improving machine lip-reading. In *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 2009–2013). doi:10.1109/ICASSP.2016.7472029.

Benezeth, Y., & Bachman, G. (2011). BL-Database: A French audiovisual database for speech driven lip animation systems. http://hal.inria.fr/inria-00614761/.

Bernstein, L. (1991). Lipreading Corpus V-VI: Disc 3 and Corpus VI-VIII: Disc 4.

Biswas, A., Sahu, P., & Chandra, M. (2015). Multiple camera in car audio–visual speech recognition using phonetic and visemic information. *Comput Electr Eng*, *47*, 35–50. doi:10.1016/j.compeleceng.2015.08.009, http://linkinghub.elsevier.com/retrieve/pii/S0045790615002864.

Bolia, R.S., Nelson, W.T., Ma, E., & Simpson, B.D. (2000). A speech corpus for multitalker communications research. *J Acoust Soc Amer*, *107*(2), 1065–1066. doi:10.1121/1.428288.

Bratoszewski, P., Lopatka, K., & Czyzewski, A. (2014). Examining Influence Of Video Framerate And Audio / Video Synchronization On Audio-Visual Speech Recognition Accuracy. In *15th International Symposium on New Trends in Audio and Video* (pp. 25–27): Wroclaw, Poland.

Bratoszewski, P., Szykulski, M., & Czyzewski, A. (2015). Examining influence of distance to microphone on accuracy of speech recognition. In *Audio Engineering Society Convention 138,* http://www.aes.org/e-lib/browse.cfm?elib=17629.

Chibelushi, C.C., Gandon, S., Mason, J.S.D., Deravi, F., & Johnston, R.D. (1996). Design issues for a digital audio-visual integrated database. doi:10.1049/ic:19961151.

Chibelushi, C.C., Deravi, F., & Mason, J.S.D. (2002). A review of speech-based bimodal recognition. doi:10.1109/6046.985551.

Chitu, A.G., & Rothkrantz, L.J.M. (2007). Building a data corpus for audio-visual speech recognition. *Euromedia '2007*, *1*(Movellan 1995), 88–92. URL <Go to ISI>://WOS:000255591600012.

Chung, J.S., Senior, A., Vinyals, O., & Zisserman, A. (2016). Lip reading sentences in the wild. In arXiv:1611.05358.

Cooke, M., Barker, J., Cunningham, S., & Shao, X. (2006). An audio-visual corpus for speech perception and automatic speech recognition. *J Acoust Soc Amer*, *120*(5 Pt 1), 2421–2424. doi:10.1121/1.2229005.

Czyzewski, A., Kaczmarek, A., & Kostek, B. (2003). Intelligent processing of stuttered speech. *J Intell Inf Syst*, *21*(2), 143–171. doi:10.1023/A.1024710532716.

Czyzewski, A., Kostek, B., Ciszewski, T., & Majewicz, D. (2013). Language material for english audiovisual speech recognition system development. *Proc Meet Acoust*, *20*(1), 060002. doi:10.1121/1.4864363.

Dalka, P., Bratoszewski, P., & Czyzewski, A. (2014). Visual lip contour detection for the purpose of speech recognition. doi:10.1109/ICSES.2014.6948716.

Davis, S., & Mermelstein, P. (1980). Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. doi:10.1109/TASSP.1980.1163420.

Durand, J., Gut, U., & Kristoffersen, G. (2014). The oxford handbook of corpus phonology. doi:10.1093/oxfordhb/9780199571932.

Fox, N.A., O'Mullane, B.A., & Reilly, R.B. (2005). VALID: A new practical audio-visual database, and comparative results. Audio-and Video-Based Biometric Person Authentication pp 777–786, doi:10.1007/11527923_81.

Garofolo, J., Lamel, L., Fisher, W., Fiscus, J., Pallet, D., Dahlgren, N., & Zue, V. (1993). TIMIT Acoustic-Phonetic Continuous Speech Corpus LDC93S1. Web Download.

Igras, M., Ziółko, B., & Jadczy, T. (2012). Audiovisual database of polish speech recordings. *Stud Inf*, *33*(2B), 163–172. doi:10.5072/si2012_v33.n2B.182.

Jadczyk, T., & Ziółko, M. (2015). Audio-visual speech processing system for polish with dynamic bayesian network models. In *Proceedings of the World Congress on Electrical Engineering and Computer Systems and Science,* http://avestia.com/EECSS2015_Proceedings/files/papers/MVML343.pdf.

Kashiwagi, Y., Suzuki, M., Minematsu, N., & Hirose, K. (2012). Audio-visual feature integration based on piecewise linear transformation for noise robust automatic speech recognition. doi:10.1109/SLT.2012.6424213.

Kunka, B., Kupryjanow, A., Dalka, P., Bratoszewski, P., Szczodrak, M., Spaleniak, P., Szykulski, M., & Czyzewski, A. (2013). Multimodal English corpus for automatic speech recognition. In *Signal Processing - Algorithms, Architectures, Arrangements, and Applications Conference Proceedings, SPA, pp. 106–111,* http://www.scopus.com/inward/record.url?eid=2-s2.0-84897901272&partnerID=tZOtx3y1.

Kuwabara, H. (1996). Acoustic properties of phonemes in continuous speech for different speaking rate. doi:10.1109/ICSLP.1996.607301.

Lane, H., & Tranel, B. (1971). The lombard sign and the role of hearing in speech. *J Speech Lang, Hear Res*, *14*, 677–709. doi:10.1044/jshr1404.677.

Lane, H., & Tranel, B. (1993). The lombard reflex and its role on human listeners and automatic speech recognizers. *J Acoust Soc Amer*, *93*, 510–524. doi:10.1044/jshr.1404.677.

Lee, B., Hasegawa-johnson, M., Goudeseune, C., Kamdar, S., Borys, S., Liu, M., & Huang, T. (2004). AVICAR : Audio-Visual Speech Corpus in a Car Environment. 8th International Conference on Spoken Language Processing pp 8–11.

Lopatka, K., Kotus, J., Bratoszewski, P., Spaleniak, P., Szykulski, M., & Czyzewski, A. (2015). Enhanced voice user interface employing spatial filtration of signals from acoustic vector sensor. Proceedings - 2015 8th International Conference on Human System Interaction, HSI 2015 pp 82–87, doi:10.1109/HSI.2015.7170647.

McCool, C., Marcel, S., Hadid, A., Pietikainen, M., Matejka, P., Cernock, J., Poh, N., Kittler, J., Larcher, A., Levy, C., Matrouf, D., Bonastre, J.F., Tresadern, P., & Cootes, T. (2012). Bi-modal person recognition on a mobile phone: Using mobile phone data. Proceedings of the 2012 IEEE International Conference on Multimedia and Expo Workshops, ICMEW 2012 pp 635–640, doi:10.1109/ICMEW.2012.116.

McGurk, H., & MacDonald, J. (1976). Hearing lips and seeing voices. *Nature*, *264*(5588), 746–748. doi:10.1038/264746a0.

Messer, K., Matas, J., Kittler, J., & Jonsson, K. (1999). XM2VTSDB: The extended M2VTS database. In *Second International Conference on Audio and Video-based Biometric Person Authentication* (pp. 72–77).

Movellan, J.R. (1995). Visual speech recognition with stochastic networks, In Tesauro, G., Touretzky, D.S., & Leen, T.K. (Eds.) *Advances in Neural Information Processing Systems 7* (pp. 851–858): MIT Press. http://papers.nips.cc/paper/993-visual-speech-recognition-with-stochastic-networks.pdf.

Mroueh, Y., Marcheret, E., & Goel, V. (2015). Deep multimodal learning for audio-visual speech recognition. In *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 2130–2134). doi:10.1109/ICASSP.2015.7178347.

Nguyen, Q.D., & Milgram, M. (2009). Semi Adaptive Appearance Models for lip tracking. doi:10.1109/ICIP.2009.5414105.

Noda, K., Yamaguchi, Y., Nakadai, K., Okuno, H.G., & Ogata, T. (2015). Audio-visual speech recognition using deep learning. *Appl Intell*, *42*(4), 722–737. doi:10.1007/s10489-014-0629-7.

Park, Y., Patwardhan, S., Visweswariah, K., & Gates, S.C. (2008). An Empirical Analysis of Word Error Rate and Keyword Error Rate.

Patterson, E.K., Gurbuz, S., Tufekci, Z., & Gowdy, J.N. (2002). CUAVE: A new audio-visual database for multimodal human-computer interface research. doi:10.1109/ICASSP.2002.5745028.

Petajan, E.D., Bischoff, B., & Bodoff, D. (1988). An improved automatic lipreading system to enhance speech recognition. Human Factors in Computing Systems Conference pp 19–25, doi:10.1145/57167.57170.

Petrovska-Delacrétaz, D., Lelandais, S., Colineau, J., Chen, L., Dorizzi, B., Ardabilian, M., Krichen, E., Mellakh, M.A., Chaari, A., Guerfi, S., D'Hose, J., & Amor, B.B. (2008). The IV2 multimodal biometric database (including Iris, 2D, 3D, stereoscopic, and talking face data), and the IV2-2007 evaluation campaign. *BTAS 2008 - IEEE 2nd Int Conf Biom: Theory, Appl Syst*, *00*, 3–9. doi:10.1109/BTAS.2008.4699323.

Pigeon, S., & Vandendorpe, L. (1997). The M2VTS multimodal face database (Release 1.00). *Audio-Video-based Biom Person Authentication*, *1206*, 403–409. doi:10.1007/BFb0015972, 10.1007/BFb0016021.

Potamianos, G., Neti, C., & Deligne, S. (2003). *Joint audio-visual speech processing for recognition and enhancement*.

Sanderson, C., & Lovell, B.C. (2009). Advances in Biometrics: Third International Conference, ICB 2009, Alghero, Italy, 2009. Proceedings, Springer Berlin Heidelberg, Berlin, Heidelberg, chap Multi-Regi, pp 199–208. doi:10.1007/978-3-642-01793-3_21.

Stewart, D., Seymour, R., Pass, A., & Ming, J. (2014). Robust audio-visual speech recognition under noisy audio-video conditions. *IEEE Trans Cybern*, *44*(2), 175–184. doi:10.1109/TCYB.2013.2250954.

Teferi, D., & Bigun, J. (2008). Evaluation protocol for the DXM2VTS database and performance comparison of face detection and face tracking on video. doi:10.1109/ICPR.2008.4761875.

Trentin, E., & Matassoni, M. (2003). Noise-tolerant speech recognition: The SNN-TA approach. *Inf Sci*, *156*(1–2), 55–69. doi:10.1016/S0020-0255(03)00164-6.

Trojanová, J., Hrúz, M., Campr, P., & Zelezny, M. (2008). Design and Recording of Czech Audio-Visual Database with Impaired Conditions for Continuous Speech Recognition. Proceedings of the Sixth International Language Resources and Evaluation (LREC'08) pp 1239–1243, http://www.lrec-conf.org/proceedings/lrec2008/.

Vlaj, D., & Kacic, Z. (2011). Computer Science and Engineering doi:10.5772/17520, http://www.intechopen.com/books/speech-technologies/the-influence-of-lombard-effect-on-speech-recognition.

Vorwerk A., Wang X., Kolossa D., Zeiler S., & Orglmeister R. (2010). WAPUSK20 - A database for robust audiovisual speech recognition, In Calzolari, N., Choukri, K., Maegaard, B., Mariani, J., Odijk, J., Piperidis, S., Rosner, M., & Tapias, D. (Eds.) *LREC, European Language Resources Association,* http://dblp.uni-trier.de/db/conf/lrec/lrec2010.html#VorwerkWKZO10.

Wong, Y.W., Ch'Ng, S.I., Seng, K.P., Ang, L.M., Chin, S.W., Chew, W.J., & Lim, K.H. (2011). A new multi-purpose audio-visual UNMC-VIER database with multiple variabilities. *Pattern Recogn Lett*, *32*(13), 1503–1510. doi:10.1016/j.patrec.2011.06.011.

Young, S.J., Kershaw, D., Odell, J., Ollason, D., Valtchev, V., & Woodland, P. (2006). *The HTK Book Version 3.4*: Cambridge University Press.

Żelasko, P., Ziółko, B., Jadczyk, T., & Skurzok, D. (2016). Agh corpus of polish speech. *Lang Resour Eval*, *50*(3), 585–601. doi:10.1007/s10579-015-9302-y.