AVICAR: Audio-Visual Speech Corpus in a Car Environment

Bowon Lee, Mark Hasegawa-Johnson, Camille Goudeseune,
Suketu Kamdar, Sarah Borys, Ming Liu, Thomas Huang

Department of Electrical and Computer Engineering
University of Illinois at Urbana-Champaign, Urbana, IL
{bowonlee, jhasegaw, cog, skamdar, sborys, minglui1}@uiuc.edu, huang@ifp.uiuc.edu

Abstract

We describe a large audio-visual speech corpus recorded in a car environment, as well as the equipment and procedures used to build this corpus. Data are collected through a multi-sensory array consisting of eight microphones on the sun visor and four video cameras on the dashboard. The script for the corpus consists of four categories: isolated digits, isolated letters, phone numbers, and sentences, all in English. Speakers from various language backgrounds are included, 50 male and 50 female.

In order to vary the signal-to-noise ratio, each script has five different noise conditions: idling, driving at 35 mph with windows open and closed, and driving at 55 mph with windows open and closed. The corpus is available through <http://www.ifp.uiuc.edu/speech/AVICAR/>.

1. Introduction

Human perception of speech is a multimodal process. The acoustic speech signal is the primary cue for recognizing speech, but visual observation of the lips, teeth, tongue, and jaw contribute to perception of phoneme articulation, while the angle of the head and raising of the eyebrows help convey sentence-level prosody [1]. Human behavior of combining audio and visual information for speech recognition is well demonstrated by the McGurk effect [2] in which the discrepancy between audio and visual information results in perceptual confusion. Visual information plays an important role especially in noisy environments [3] [4] which encourage the additional use of visual information to increase speech recognition accuracy.

Automatic speech recognition (ASR) achieves higher than 99% correct recognition accuracy for connected digits using a hidden Markov Model (HMM) recognizer specifically designed for this task [5]. However, the performance degrades severely when the training and test data sets have mismatched noise conditions such as signal-to-noise ratio (SNR) or speaking styles. To build a robust speech recognizer, a very large training data set may be used to cover all possible types of acoustic variability. Feature and model compensation can be used to reduce the sensitivity of an HMM to acoustic noise [6], but even noise-compensated models perform best when trained using mixed-SNR data [7].

Background noise affects the acoustic environment as an additive noise signal at the microphone, but more importantly by causing the speaker to increase vocal effort to overcome noise levels in his own ears (the Lombard effect) [8]. The variation of speech production caused by noise exposure at the ear can degrade performance more than the ambient noise itself [9]. Because of this, simply adding additive noise to speech data recorded in a quiet environment may not produce training data that adequately represents real-world test conditions.

One of the noisy environments in need of a speech recognizer is the inside of an automobile. Operating certain devices such as a telephone or a car navigation system by hand may distract the driver. Even basic functions such as operating an air conditioner or the car audio system may cause undesirable distraction. A reliable speech recognition system for the automobile environment can be a good alternative to manual operation of those functions.

In our research, we have found that typical acoustic background noise levels in a car vary from approximately 15 dB SNR to -10 dB SNR. For automobile environments, various microphone types and positions can increase SNR [10]. Another approach is to use a microphone array to increase the SNR [11]. In order to try methods to improve the performance of speech recognizers for the automobile environment, it is necessary to have an extensive speech database recorded in actual cars. SpeechDat-Car [12] is a multilingual database of nine European languages recorded in an automobile environment. It started in 1998 and 600 sessions will be recorded with at least 300 speakers for each language. It uses a total of four microphones, one near-field microphone as a reference and three far-field microphones. CU-Move [13] is a database using an array of five microphones with a reference microphone to capture background noise. An overview of various car speech databases is presented in [14].

Combining visual and audio information can improve ASR accuracy for low SNR conditions [15]. For humans, it has been shown that the presence of the visual signal is roughly equivalent to a 12 dB gain in acoustic SNR [3]. Automatic systems can show similar benefits: the combination of audio and visual features using a coupled HMM (CHMM) can improve word recognition accuracy by more than 40% at 20 dB SNR with additive white Gaussisan noise [16].

Audio-visual databases currently available include MOCHA [17] and CUAVE [18]. Because of the large size of each data file, the number of speakers is limited: 10 for MOCHA and 37 for CUAVE. The size of vocabulary is also limited: 78 isolated words for MOCHA and only connected and isolated digits for CUAVE. These databases are recorded in a quiet office environment with only one microphone.

We are interested in low SNR speech recognition using a microphone array combined with visual information to increase accuracy. In order to facilitate study of this problem, we have collected a speech recognition training and test database recorded in a moving automobile using an array of four cameras and eight microphones (AVICAR: *a*udio-*v*isual speech *i*n a *car*).

2. Array of sensors

2.1. Microphone array and beamforming

The purpose of using an array of microphones as input to an ASR system is its ability to acquire an enhanced signal using beamforming algorithms. Delay-and-sum beamforming improves SNR by suppressing sources off the main axis of the beam. Adaptive beamformers selectively suppress the noise power incident from directions other than the source [19] [20]. Various microphone array processing methods are well summarized in [11].

Automobile environments have various kinds of noise such as wind noise, road noise, and vehicles passing by. These noise sources originate from different directions than the speaker, so their impact can be minimized by using a microphone array. It has been demonstrated that beamforming can reduce the word error rate of a speech recognizer in noisy environments [21] [22].

2.2. Visual feature extraction and 3D face modeling

An array of cameras allows for the extraction of 3D shape-based features for audio-visual speech recognition. The main approaches for visual feature extraction from image sequences can be grouped into image-based, visual-motion-based, geometric-feature-based, and model-based approaches [23]. The advantage of this last approach is that model-based features can often be made invariant to image transformations such as translation, rotation, and lighting [23]. One problem with the video data captured in an automobile is an illumination effect: lighting conditions vary widely and these changing light conditions are likely to dominate the observed distribution of image-based audio-visual speech recognition features, substantially degrading the word recognition accuracy at low SNR. To compensate for variable lighting, we propose to supplement image-based features with 3D model-based features extracted from a camera array. 3D models of lip movement can be estimated from 2D image data [24]. It is possible to construct a 3D model and facial feature extraction with images from different angles [25].

3. Equipment

3.1. Audio

An array of eight omnidirectional microphones captures audio. The off-the-shelf cell phone microphones and the LM386 audio preamplifiers are inexpensive and comparable to what a commercial system would use. Each microphone is 6 mm in diameter. They are spaced 1.5 inches apart. Microphone preamplifiers are mounted at the microphones on the sun visor. Seven of the eight preamplified audio channels are sent to an ADAT (Advanced Digital Audio Tape) through shielded cables. The ADAT records eight audio channels at 16 bit resolution with a sampling rate of 48 kHz. We considered recording audio directly to the hard disk of a laptop, but the ADAT tapes provide a safe backup, a "camera master" in the jargon of professional video production. One channel of the ADAT is reserved for the control sequence, described in section 3.3.

3.2. Video

Cameras above the windshield or far to the side have a poor view of the subject's mouth, so the cameras are placed on the dashboard. Mounting more than four cameras in this limited space does not increase the amount of useful data enough to warrant the extra complexity of more camcorders and more data files. Therefore, we chose an array of four cameras to capture video (Fig. 1). Each camera is aimed from different positions on the dashboard to capture the face region of a person sitting in the front passenger seat. Black cardboard lens hoods reduce glare and also help in aiming the cameras. The four video streams are combined by a video multiplexer and sent to a MiniDV camcorder. One of the two audio channels of the camcorder is used for the eighth microphone input from the microphone array; the other is used for a control sequence which is exactly the same as that recorded by the ADAT. The camcorder uses the same audio resolution and sampling rate as the ADAT.

Figure 1: *Eight microphones on the sun visor, four cameras on the dashboard.*

3.3. Control sequence

Since the corpus is multimodal, we need to synchronize the data. Also, we need to segment the data from the raw recordings into individual utterance units. For efficiency in building a database, automatic synchronization and segmentation is required.

We use DTMF (Dual Tone Multi-Frequency) tones as a control sequence. DTMF control tones are generated by a telephone handset held by the subject (Fig. 2). As with most of the other equipment, the telephone handset is inexpensive consumer technology: tested to survive abuse, and easily replaceable when it does fail. When back in the laboratory, we use similarly robust DTMF detection software.

One tone out of ten digits is assigned to each utterance in the script, and subjects are asked to press the assigned button before speaking. Mistakes are marked with the '$*$' button, pauses with the '$\#$' button. The DTMF signal is fed into both the ADAT and camcorder.

During postprocessing, segmentation and labeling are done with the information about which tone is detected. The onset of each tone synchronizes the recordings from the ADAT and camcorder.

3.4. Installation

For the equipment, we have three signal sources (array of microphones, cameras, and DTMF generator) and two recording devices (ADAT and MiniDV camcorder). Recording devices are located in the back seat. Every device is powered by the AC power from a DC to AC converter except for the 9 V

battery-powered microphone preamplifiers. Equipment installation takes one person 20 minutes, removal 10 minutes. Figure 3 shows all the signal paths.

4. Speakers and scripts

The corpus includes 100 speakers, 50 male and 50 female. About 60% are native speakers of American English, while others have Latin American, European, East Asian, and South Asian backgrounds. All speech recorded for the database is in English.

Table 1: *Categories for each script set.*

Category	Examples
Isolated Digits	one, two, · · · , ten, oh, zero, done
Isolated Letters	a, b, c, · · · , z
Phone numbers	(163)516-3885
TIMIT Sentences	When all else fails, use force.

Ten different script sets are used for the corpus and each set is for ten speakers, five male and five female. Categories of each script set are listed in Table 1. Isolated digits are used to train recognition models for automatic dialing purposes. Isolated letters are useful for the study of difficult phonetic contrasts, e.g., 'bee' vs. 'dee.' Phone numbers are included as connected digits because automatic dialing is a potentially important application of this technology. Phonetically balanced TIMIT sentences are included to provide training and test data for phoneme-based recognizers [26]. Subjects are asked to speak isolated digits and letters twice under each noise condition. Each script set has 20 phone numbers with 10 digits each, a total of 200 individual digits chosen randomly with uniform frequency. For 10 phone numbers '0' is pronounced as 'zero', for the other 10 phone numbers, 'oh.' A total of 20 sentences are used in each script set. These sentences are randomly chosen out of 450 phonetically compact sentences from the TIMIT speech database. Each speaker reads one script set repeatedly under five different noise conditions. The vocabulary of this corpus consists of 13 digits, 26 isolated letters, and other words in TIMIT sentences, for a total vocabulary size of 1,317 words. The total number of utterances is 118 for each script set. Since these scripts are recorded from 100 speakers and repeated under five different noise conditions, the total number of utterances is 59,000 recorded in eight audio and four video channels.

Figure 2: *Script with the telephone handset.*

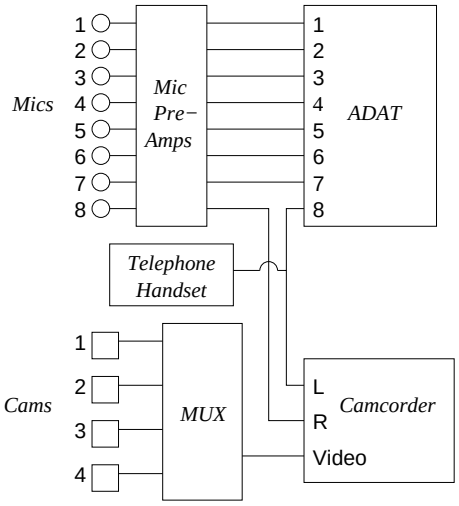

Figure 3: *Equipment setup.*

5. Post processing

Audio and video data are collected on the ADAT and MiniDV tapes during the recording session. Data on those magnetic tapes are transferred to the computer for post processing.

5.1. Audio

Audio data in the ADAT are transferred via an optical cable through a multi-channel Firewire (IEEE 1394) audio interface (MOTU 828mkII) to the computer. Eight-channel audio signals are saved as .WAV files. The eighth channel which contains the DTMF control sequence is separated for onset detection and identification of each tone. Audio data are downsampled from 48 kHz to 16 kHz after segmentation. Speech amplitude varies as a function of noise level because of the Lombard effect (Fig. 4).

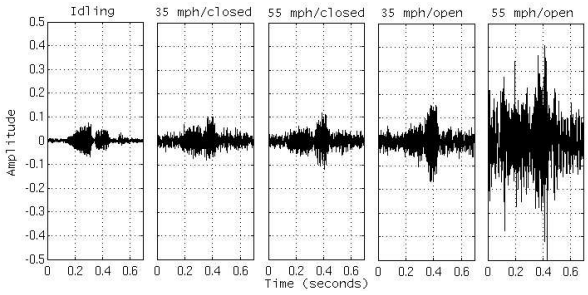

Figure 4: *The speech waveform 'seven' under five noise conditions.*

5.2. Video

Video data including the two-channel audio data in MiniDV tape are transferred through the Firewire interface to the computer. Audio channel containing the control sequence is separated for segmentation. One channel containing the eighth microphone signal is segmented according to the control sequence and added to the audio data. The video stream (Fig. 5) is encoded to a compressed format to reduce the size and then also segmented according to the control sequence.

Figure 5: *A snapshot of the video stream.*

6. Conclusions

We have built a speaker-independent multi-sensory audio-visual speech corpus in a car environment. This database is available by request from <http://www.ifp.uiuc.edu/speech/AVICAR/>.

7. Acknowledgement

The authors would like to thank the Integrated Systems Laboratory, Beckman Institute, for hardware assistance. This project is funded by Motorola Research, Schaumburg, IL.

8. References

[1] B. Granström and D. House, "Audiovisual representation of prosody in expressive speech communication," *ISCA Int. Conf. Speech Prosody*, pp. 393–400, 2004.

[2] H. McGurk and J. MacDonald, "Hearing lips and seeing voices," *Nature*, vol. 264, pp. 746–748, 1976.

[3] W. H. Sumby and I. Pollak, "Visual contributions to speech intelligibility in noise," *J. Acoust. Soc. Am.*, vol. 26, no. 2, pp. 212–215, 1954.

[4] K. W. Grant and L. D. Braida, "Evaluating the articulation index for auditory-visual input," *J. Acoust. Soc. Am.*, vol. 89, no. 6, pp. 2952–2960, 1991.

[5] W. Chou, C.-H. Lee, and B. H. Juang, "Minimum error rate training of inter-word context dependent acoustic model units in speech recognition," *Proc. Int. Conf. Spoken Lang. Process.*, pp. 439–442, 1994.

[6] C.-H. Lee, "On stochastic feature and model compensation approaches to robust speech recognition," *Speech Comm.*, vol. 25, no. 1, pp. 29–47, 1998.

[7] M. Matassoni, M. Omologo, and P. Svaizer, "Use of real and contaminated speech for training of a hands-free in-car speech recognizer," *Eurospeech*, 2001.

[8] E. Lombard, "Le signe de l'elevation de la voix," *Ann. Maladies Oreille, Larynx, Nez, Pharynx*, vol. 37, pp. 101–119, 1911.

[9] P. Rajasekaran, G. Doddington, and J. Picone, "Recognition of speech under stress and in noise," *Proc. Int. Conf. Acoust., Speech, and Sig. Process.*, pp. 733–736, 1986.

[10] R. Aubauer and D. Leckschat, "Optimized second-order gradient microphone for hands-free speech recordings in cars," *Speech Comm.*, vol. 34, no. 1-2, pp. 13–23, 2001.

[11] M. S. Brandstein and D. B. Ward, *Microphone Arrays: Signal Processing Techniques and Applications*. Springer Verlag, 2001.

[12] H. V. den Heuvel, R. Boudy, S. Euler, A. Moreno, and G. Richard, "The SpeechDat-Car multilingual speech databases for in-car applications: Some first validation results," *Eurospeech*, pp. 2279–2282, 1999.

[13] J. H. L. Hansen, J. Plucienkowski, S. Gallant, B. Pellom, and W. Ward, "CU-Move: Robust speech processing for in-vehicle speech systems," *Proc. Int. Conf. Spoken Lang. Process.*, pp. 524–527, 2000.

[14] D. Langmann, H. R. Pfitzinger, T. Schneider, R. Grudszus, A. Fischer, M. Westphal, T. Crull, and U. Jekosch, "CSDC – the MoTiV car speech data collection," *Proc. Int. Conf. Lang. Resources and Eval.*, pp. 1107–1110, 1998.

[15] T. Chen, "Audiovisual speech processing," *IEEE Sig. Process. Magazine*, vol. 18, no. 1, pp. 9–21, 2001.

[16] S. Chu and T. Huang, "Audio-visual speech modeling using coupled hidden Markov models," *Proc. Int. Conf. Acoust., Speech, and Sig. Process.*, pp. 2009–2012, 2002.

[17] http://amp.ece.cmu.edu/projects/AudioVisualSpeechProcessing/.

[18] E. K. Patterson, S. Gurbuz, Z. Tufekci, and J. N. Gowdy, "CUAVE: A new audio-visual database for multimodal human-computer interface research," *Proc. Int. Conf. Acoust., Speech, and Sig. Process.*, pp. 2017–2020, 2002.

[19] O. L. Frost, III, "An algorithm for linearly constrained adaptive array processing," *Proc. of IEEE*, vol. 60, no. 8, pp. 926–935, 1972.

[20] L. J. Griffiths and C. W. Jim, "An alternative approach to linearly constrained adaptive beamforming," *IEEE Trans. Antennas and Propag.*, vol. 30, no. 1, pp. 27–34, 1982.

[21] J. Bitzer, K. U. Simmer, and K.-D. Kammeyer, "Multi-microphone noise reduction techniques as front-end devices for speech recognition," *Speech Comm.*, vol. 34, pp. 3–12, 2001.

[22] T. B. Hughes, H.-S. Kim, J. H. DiBiase, and H. F. Silverman, "Performance of an hmm speech recognizer using a real-time tracking microphone array as input," *IEEE Trans. Speech and Audio Process.*, vol. 7, no. 3, pp. 346–349, 1999.

[23] S. Dupont and J. Luettin, "Audio-visual speech modeling for continuous speech recognition," *IEEE Trans. Multim.*, vol. 2, no. 3, pp. 141–151, 2000.

[24] S. Basu, N. Oliver, and A. Pentland, "3D modeling and tracking of human lip motions," *Proc. Sixth Int. Conf. Computer Vision*, pp. 337–343, 1998.

[25] F. Pighin, R. Szeliski, and D. H. Salesin, "Modeling and animating realistic faces from images," *Int. J. of Computer Vision*, vol. 50, no. 2, pp. 143–169, 2002.

[26] V. Zue, S. Seneff, and J. Glass, "Speech database development at MIT: TIMIT and beyond," *Speech Comm.*, vol. 9, no. 4, pp. 351–356, 1990.