ISCA Archive
http://www.isca-speech.org/archive

INTERSPEECH 2012
ISCA's 13th Annual Conference
Portland, OR, USA
September 9-13, 2012

The 'Audio-Visual Face Cover Corpus': Investigations into audio-visual speech and speaker recognition when the speaker's face is occluded by facewear

Natalie Fecher

Department of Language and Linguistic Science, University of York, York, United Kingdom
natalie.fecher@york.ac.uk

Abstract

The Audio-Visual Face Cover Corpus consists of high-quality audio and video recordings of 10 native British English speakers wearing different types of 'facewear'. Speakers read aloud a set of 64 /C_1VC_2/ syllables embedded in a carrier phrase. 18 English consonants occurred twice each in onset and coda positions. Speakers recited the list 1+8 times, i.e. once in control condition (no facewear) and eight times while wearing a forensically-relevant face covering. Audio recordings were made by simultaneously capturing the speech via a headband microphone and two shotgun microphones placed facing and behind the speaker. Footage of the subject's head and shoulders was filmed from two camera angles, frontal and half-profile. In total, 6,120 utterances were recorded per device. This paper aims to specify the database design, to introduce forensic-phonetic research utilising the data, and to demonstrate the corpus's potential applications in related fields of study and in casework conducted by forensic speech scientists.

Index Terms: speech database, audio-visual, forensic speech science, facewear, disguise, acoustic phonetics, perception

10.21437/Interspeech.2012-133

1. Introduction

The primary intention for the collection of the Audio-Visual Face Cover Corpus (AVFCC) was to provide multi-purpose audio-visual (AV) speech data comprised of speakers wearing different types of 'facewear'. The corpus aims to facilitate research on verbal communication situations that can occur in very specific forensic scenarios, namely ones in which the speaker's face is disguised by some sort of face-concealing garment or headgear. The high-quality AV speech recordings collected provide the basic data set for a project being carried out within the Marie Curie Initial Training Network 'Bayesian Biometrics for Forensics (BBfor2)' and which investigates the forensic implications of multimodal speech and speaker recognition. It mainly addresses the acoustic-phonetic and perceptual impact of face coverings on speech intelligibility and voice identification, expanding research in this underexamined area of forensic (speech) science [1]. Forensic speech science is a highly interdisciplinary field that applies and extends expertise and methods from, among others, phonetics, (socio)linguistics and acoustic signal processing to practical tasks arising in the context of police work or the presentation of evidence in court [2]. The AVFCC corpus can provide reference material for forensic-phonetic and acoustic work on authentic cases involving speakers whose facial appearance is fully or partially disguised. If evidence becomes available that the speech in dispute in an investigation was produced through facewear, the forensic expert should be prepared to take that knowledge into account as a potentially influencing factor in his/her auditory, acoustic and (where applicable) automatic analysis of the material. Such evidence may emerge in the form of circumstantial information provided e.g. by the police, or via indicators present in the speech sample itself (e.g. a recording of an emergency call during which the talker's mouth was apparently taped shut).

In addition to its practical relevance to the forensic community the AVFCC corpus is intended as a resource for empirical studies on *audio-only* speech processing as well as *audio-visual* and *visual-only* (lipreading) speech and face processing. The data were collected in acknowledgement of the fact that the occlusion of a speaker's face while speaking is likely to have a combined acoustic, auditory and visual impact on speech and speaker recognition [1, 3]. Human perception of the environment, including spoken language perception, is multi-sensory and multi-modal [4]. Numerous studies, such as [5-11], have demonstrated that both acoustic and visual signals generated during speech production play a fundamental role in the perceiver's recognition of speech content and/or speaker identity. In recent years, several multimodal databases have been created for the purpose of testing recognition performance in automatic multibiometric recognition systems [12]. Often, however, the speech material is phonetically and acoustically unsuitable for perceptual testing with human subjects, especially where the focus is on speech perception in adverse listening and/or viewing conditions [13, 14]. To examine the effects of signal degradation of this kind, additive audio and image noise or channel distortions are frequently deployed; another means of impeding speech processing and interpretation is by occluding the speaker's face. AV speech recognition during facial occlusion is tested in several studies by post-processing the visual (speech) image, for instance by blurring or blacking out carefully selected parts of the face (see e.g. [5, 8]). However, this method was not considered appropriate for the 'real-life' forensic context under investigation here, bearing in mind the assertion made in [1] and [3] that the impact of facewear on the acoustic signal is likely to stem from two different sources. These sources are the acoustic impedance characteristics of the mask material (see §2.2) and modifications to the output signal brought about by the facewear's interaction with the speech articulators. As is known from acoustic-phonetic theory, even minor repositioning of the speaker's articulators can give rise to prominent acoustic and perceptual changes [15]. To sum up, only very few corpora adopt facial occlusion as a within-subject design parameter, e.g. the M2VTS or BT-DAVID corpus (spectacles/hats/scarves; see [4, 16]). The AVFCC corpus is the first of its kind as it includes a considerable variety of facewear. To increase reusability of the data and to compensate for the small set of speakers, the corpus design (see §2) incorporated different microphone positions, camera angles, and the option for chroma-keying (compositing technique for replacing a monochromatic background of a moving or still image with a different image in post-production).

2. Corpus design

2.1. Speakers

The AVFCC corpus consists of recordings of ten speakers, five females and five males. Their ages ranged from 21-36 (mean: 26.5, SD: 5.7). No participant reported a history of impaired speech, hearing or vision. All were native English speakers who talk generally with a Southern Standard British English accent. All of them had a linguistics/phonetics background and held a degree in linguistics from BA to PhD level at the date of recording. Hence, all speakers had previous training in the International Phonetic Alphabet (IPA), which enabled them to produce the target stimuli presented using IPA characters reliably. No participant reported prior experience of wearing any type of facewear for recreational, occupational or religious purposes on a regular basis. Given the variety of facewear tested, it seemed more feasible to recruit speakers with limited experience of wearing facewear. This factor was controlled for, as people who routinely wear a face covering, e.g. surgical nurses or doctors, may compensate for known disadvantageous auditory effects more extensively, e.g. by applying production/ articulation compensation strategies such as raised voice or hyperarticulation. Speakers were staff and students recruited at the Department of Language and Linguistic Science, York, and were paid for their participation.

2.2. Facewear

The term 'facewear' is introduced in this context to refer to various types of face-concealing garments and headgear worn, on the one hand, for the commission of crimes such as armed robberies, assaults, or terrorist and paramilitary activities. Figure 1 shows images of one of the corpus speakers wearing examples of these types of masks, e.g. a balaclava, hoodie/scarf combination, or motorcycle crash helmet.

Figure 1: *Speaker in control condition and wearing eight types of facewear. Selection criteria for facewear were forensic relevance, region of facial occlusion and facewear material.*

The corpus, on the other hand, includes facewear worn for occupational and recreational reasons, e.g. a surgical mask or motorcycle helmet, as well as for religious reasons (a niqāb, i.e. a Muslim full-face veil). This shows that the selection of relevant face coverings was not solely motivated by their direct forensic relevance, but was also targeted at everyday spoken interactions out of which a forensic case may potentially arise. (For arguments based on intelligibility in the context of the recent 'burka' discussions across Europe see [1].)

A second selection criterion besides forensic relevance was to do with the parts of the speaker's face concealed, i.e. if the mouth only, or the mouth and nose, or if additionally the ears were covered (potentially provoking the Lombard reflex). The third selection criterion concerned the material that covers the speaker's articulators. This was considered a crucial factor because different materials are assumed to absorb the sound energy at different levels and in different frequency bands [1, 3, 17-19].

Finally, another factor to be kept in mind is that all speakers wore the selfsame facewear during the individual recording sessions. Naturally, these fit the speakers to varying degrees, depending on the size and shape of their heads, and may thus have perturbed articulation to a smaller or larger extent.

2.3. Speech material

Prior to reciting the main target stimuli each speaker read the 'wolf passage' [20], which obtained some phonetically-controlled reference material for each speaker and aimed to reduce speakers' stress levels at the outset of the experiment, i.e. for them to get used to the experimental setup in a large professional recording studio. The list of target stimuli was specifically designed for this corpus. It consists of phonetically controlled /C_1VC_2/ syllables. They are embedded phrase-finally in the carrier sentence *He said* <*stimulus*>. This sentence was presented to the participants as orthographic strings, whereas the syllables were displayed as IPA characters (to circumvent orthographic ambiguity). The syllables are made up of two repetitions of 18 consonants in two syllable positions. The nucleus is always the open back vowel /ɑ:/. 18 English consonants /p t k b d g f s ʃ θ v z ʒ ð m n ŋ h/ occur twice in initial and final syllable position, respectively, each time in a different phonetic environment, i.e. with different 'filler consonants'. These compensate for possible connected speech processes such as anticipatory or carryover nasal coarticulation. Phonotactic constraints were observed: /h/ in coda and /ŋ/ in onset were excluded and occurred only once apiece, making a total of 64 syllables per list. Stimuli are logatoms to prevent top-down processing, such as lexical predictability, from biasing recognition performance in subsequent perception experiments. To eliminate the number of real words in the stimuli set all tokens were checked by native English speakers. Existing one-syllable words were replaced by changing the filler consonant. In total, 6,120 utterances were recorded per device: 10 speakers * 1+8 facewear conditions * 18 consonants * 2 syllable positions * 2 repetitions (including only phonotactically-valid syllables, as stated above). To enlarge the range of potential applications of the speech data, the corpus could be extended in the future to include a larger vowel inventory, varying prosodic contexts (e.g. no phrase-final target syllables), and ideally forensically-relevant speaking styles such as emotional speech (see e.g. [23]).

2.4. Prompting method

The order of the 64 syllables in the stimuli list was randomised nine times, and each speaker read the nine lists in random order. One list was read in control condition (without facewear) and the

remaining eight with the talker each time wearing one of the face coverings, again in randomised order. The prompting method was screen-prompting; lists were presented in timed PowerPoint presentations on an LCD monitor. One stimulus sentence, e.g. *He said [zɑ:f].*, was presented per slide for 2.55s. Between each sentence a black screen was shown for 0.55s. After each block of 16 stimuli a slightly longer break of 2.5s was given to display to the speakers how many sentences were still to be read.

2.5. Speaking style

Talkers were instructed to read the stimuli carefully yet fluently, and to control their speaking style to their best ability: normal loudness, clear but not exaggerated articulation. Speaking tempo was controlled by the PowerPoint presentation. Moreover, subjects were asked to control their facial expressions where applicable: neutral, no strong eyebrow raise, start/end of each utterance with closed lips. They were advised to continue reading the list after reading or pronunciation errors as misread stimuli would be repeated at the end of each take.

2.6. Recording set-up

The database was recorded in a professional sound-treated TV studio at the Department of Theatre, Film and Television, York. Participants were seated in front of a plain green background and asked to avoid marked head movements during the recordings. Two light sources were arranged to produce a uniform illumination across the talkers' faces. They were instructed not to wear spectacles or noticeable jewellery to avoid possible reflection caused by the spotlights, and to put on black shirts provided to them. As shown in Figure 2, three simultaneous continuous audio recordings were made (RIF WAV format, 48.0kHz, 768kbit/s, 16-bit signed integer PCM encoding). A DPA 4066 Omnidirectional Headband Microphone captured the speech at approximately 2cm from the right-hand corner of each speaker's mouth. It was taped to the facewear with black or skin-coloured adhesive tape, if necessary.

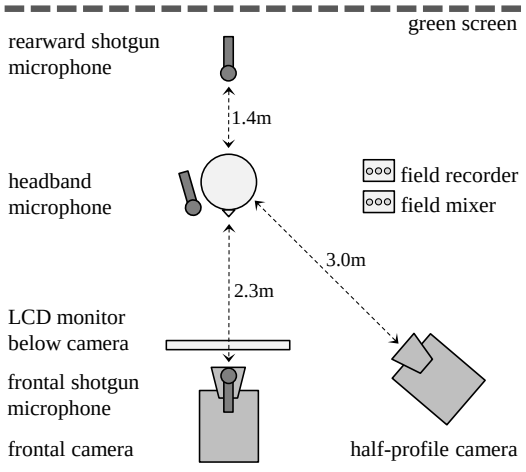

Figure 2: *The AVFCC recording set-up.*

Two Røde NTG-2 Dual Powered Shotgun Condenser Microphones captured the audio from 2.3m in front of and 1.4m behind the speaker. The rearward microphone was placed at the height of each speaker's head and is not visible in the videos. Audio was recorded with an Edirol R-4 Pro Portable 4 Channel Recorder and a Sound Devices 552 Portable Production Mixer. Two simultaneous continuous HD video recordings were made with two Panasonic AG-HPX171E Camera Recorders which were positioned so that the images consisted of the speaker's entire head and shoulders. The half-profile camera was placed opposite the location of the headband microphone so as to avoid it occluding the left-hand side of the speaker's mouth/face. The monitor for stimuli prompting was placed directly below the camera lens of the frontal camera to give the impression that the talkers were looking into the lens. The frontal camera took its audio input from the headband and the frontal microphone. The rearward microphone and the half-profile camera captured the speech separately. To facilitate the temporal alignment of all audio and video streams a clapper board signal was given at the start of each take.

3. Practical applications

3.1. Speech acoustics

As noted in §1, facewear fabrics have been shown to cause transmission loss through low-pass filtering of the speech signal. It is also assumed that facewear will affect speech production by impeding the articulators and/or the airstream. However, further research will be needed to obtain a more detailed picture of the specific nature of these 'facewear effects' on the properties of the acoustic spectrum. Similarly, the extent to which speech production and articulation is affected needs to be ascertained, for instance by applying the findings from speech perturbation studies [21]. As mentioned in §2.1, compensation strategies actively applied by the speaker as well as (passive) changes to the motor activity of the articulators due to the masks constraining natural lip/jaw movements are likely to result in changes to the acoustic signal.

3.2. Audio-visual speech perception

Absent or degraded facial speech cues provided by a talker wearing facewear may impede the perceiver's attempts to disambiguate phonemes during face-to-face communication. Several studies of bimodal speech perception (see §1) investigate the interaction of the two modalities when the image accompanying the auditory stimulus is partially or wholly obscured. There exist, however, only very few studies in which the successful recovery of crucial information from the talker's face is hindered because the face is actually concealed while s/he is talking (rather than applying a post-production mask to the video image) [22, 23]. The AVFCC corpus provides sufficiently controlled data to enable a multifaceted manipulation of both audio and video recordings, and hence the testing of human and machine performance in degraded listening conditions.

On a side note, the functionality of the experimental control software utilised in the author's ongoing perception experiments includes randomised playback of multimedia formats and stimuli control via a multiple screen interface. Please contact the author for access to the software as well as to the corpus.

3.3. Forensic-phonetic casework

An appreciable proportion of forensic-phonetic casework involves facial disguise of one form or another. However, at present we lack solid experimental data in which to ground

estimates of the influence such disguises may have on the reliability of evidence produced in connection with such cases. The quality of lay earwitness testimony (which is at best already questionable in many cases, see e.g. [24]) may be compromised further by the fact that the perceived speech was produced by a talker whose face was concealed by facewear. The witness may report being certain about the words that were used, and/or may claim that the talker's voice was that of a familiar person. Until facewear effects are better understood we cannot with any confidence say whether listeners' judgments of this kind should or can be regarded as of equivalent evidential value to those made in situations where the talker's face was not concealed. The AVFCC dataset is designed to enable research that will help to place expert opinions in this area on a firmer footing.

Similarly, the potential influence of facewear effects on the acoustic signal must be taken into account when working with speech recordings in which it is known or suspected that the talker's face was disguised by facewear. The objective which the research introduced in this paper aims to achieve is to quantify the magnitude of the changes involved. The accuracy of the observations made by expert analysts (acoustic measurements, impressionistic transcriptions, disputed utterance analysis, etc.) can only be enhanced if, in addition to all other known influencing factors, the facewear effects are taken into account (the same holds when constructing voice line-ups) [2, 24, 25, 26]. Overall, facewear should be treated as yet another factor causing inter-/intra-*speaker* as well as inter-/intra-*listener* variability [27] in the 'forensic trace', which in the present context is the audio-visual speech signal.

4. Conclusions

The Audio-Visual Face Cover Corpus, a high-quality audio and video data collection of speakers wearing various types of facewear, enables research on audio-only, audio-visual and visual-only speech and face processing. This paper specified the database design and introduced possible practical applications of the data with respect to research and casework carried out in forensic speech science. It also gave a brief overview of this comparatively novel area of forensic-phonetic research, namely speech and speaker recognition under facial disguise.

5. Acknowledgements

This project has received funding from the European Commission's Seventh Framework Programme (FP7-PEOPLE-2007) under grant agreement number 238803 (Marie Curie Initial Training Network 'Bayesian Biometrics for Forensics (BBfor2)'; http://bbfor2.net). Thanks to Dominic Watt for his valuable comments on an earlier version of this paper, as well as to Huw Llewelyn-Jones, Tai Chi Minh Ralph Eastwood, Omer Qadir, three anonymous reviewers, and to all disguised speakers.

6. References

[1] Llamas, C., Harrison, Ph., Donnelly, D., and Watt, D., "Effects of different types of face coverings on speech acoustics and intelligibility", *York Papers in Linguistics (2)*, 9, 80-104, 2009.

[2] Jessen, M., "Forensic phonetics", *Language and Linguistics Compass*, 2(4), 671-711, 2008.

[3] Fecher, N. and Watt, D., "Speaking under cover: The effect of face-concealing garments on spectral properties of fricatives", *Proc. 17th Int. Congr. of Phonetic Sci.*, Hong Kong, China, 663-666, 2011.

[4] Goecke, R., "Current trends in joint audio-video signal processing: a review", *Proc. IEEE 8th Int. Symp. on Signal Process. and Its Applicat.*, 70-73, 2005.

[5] Preminger, J. E., Lin, H.-B., Payen, M., and Levitt, H., "Selective visual masking in speechreading", *J. Speech Lang. Hear. Res.*, 41(3), 564-575, 1998.

[6] Tuomainen, J., Andersen, T. S., Tiippana, K., and Sams, M., "Audio-visual Speech Perception Is Special", *Cognition*, 96(1), B13-B22, 2005.

[7] Lidestam, B. and Beskow, J., "Visual Phonemic Ambiguity and Speechreading", *J. Speech Lang. Hear. Res.*, 49(4), 835-847, 2006.

[8] Marassa, L. K. and Lansing, C. R., "Visual Word Recognition in Two Facial Motion Conditions: Full-Face Versus Lips-Plus-Mandible", *J. Speech Lang. Hear. Res.*, 38(6), 1387-94, 1995.

[9] Rosenblum, L. D., Rachel, M. M., and Sanchez, K., "Lip-read Me Now, Hear Me Better Later: Cross-modal Transfer of Talker-familiarity Effects", *Psychol. Sci.*, 18(5), 392-396, 2007.

[10] Schwartz, J.-L., Berthommier, F., and Savariaux, Chr., "Seeing to Hear Better: Evidence for Early Audio-visual Interactions in Speech Identification", *Cognition*, 93(2), B69-B78, 2004.

[11] Sheffert, S. M. and Olson, E., "Audiovisual Speech Facilitates Voice Learning", *Perc. & Psychophysics*, 66(2), 352-362, 2004.

[12] Ross, A., "An introduction to multibiometrics", *Proc. 15th Europ. Signal Process. Conf.*, 2007.

[13] Cooke, M., Barker, J., Cunningham, S., and Shao, X., "An audio-visual corpus for speech perception and automatic speech recognition", *J. Acoust. Soc. Am.*, 120(5), 2421-24, 2006.

[14] Aleksic, P. S. and Katsaggelos, A. G., "Audio-Visual Biometrics", *Proc. of the IEEE*, 94(1), 2025-44, 2006.

[15] Stevens, K. N. and Keyser, S. J., "Quantal theory, enhancement and overlap", *J. of Phonetics*, 38, 10-19, 2010.

[16] Trojanová, J., Hrúz, M., Campr, P., and Železný, M., "Design and Recording of Czech Audio-Visual Database with Impaired Conditions for Continuous Speech Recognition", *Proc. 6th Int. Conf. on Lang. Resources and Evaluation*, 1239-43, 2008.

[17] Nute, M. E. and Slater, K., "The effect of fabric parameters on sound-transmission loss", *J. of the Textile Institute*, (64)11, 652-658, 1976.

[18] Radonovich, L. J., Yanke, R., Cheng, J., and Bender, B., "Diminished speech intelligibility associated with certain types of respirators worn by healthcare workers", *J. of Occupational and Environmental Hygiene*, 7(1), 63-70, 2010.

[19] Watt, D., Llamas, C., and Harrison, Ph., "Differences in perceived sound quality between speech recordings filtered using transmission loss spectra of selected fabrics", presented at the IAFPA 2010 Annual Conference, Trier, Germany, 2010.

[20] Deterding, D., "The North Wind versus a Wolf: short texts for the description and measurement of English pronunciation", *J. Int. Phonetic Assoc.*, (36)2, 187-196, 2006.

[21] Brunner, J., *Perturbed Speech. How compensation mechanisms can inform us about phonemic targets*. Saarbrücken, Germany: Südwestdeutscher Verlag für Hochschulschriften, 2009.

[22] Coniam, D., "The impact of wearing a face mask in a high-stakes oral examination: An exploratory post-SARS study in Hong Kong", *Language Assessment Quarterly*, 2, 235-261, 2005.

[23] Heath, A. J. and Moore, K., "Earwitness Memory: Effects of Facial Concealment on the Face Overshadowing Effect", *Int. J. of Advanced Sci. and Technology*, 33, 131-140, 2011.

[24] Yarmey, A. D., "Factors Affecting Lay Person's Identification of Speakers", in *The Oxford Handbook of Language and Law*, Tiersma, P. and Solan, L., Eds., Oxford, New York: Oxford University Press, 547-556, 2012.

[25] Foulkes, P. and French, P., "Forensic Speaker Comparison", in *The Oxford Handbook of Language and Law*, Tiersma, P. and Solan, L., Eds., Oxford, New York: Oxford University Press, 557-572, 2012.

[26] Zhang, C. and Tan, T., "Voice disguise and automatic speaker recognition", *Forensic Sci. Int.*, (175)2-3, 118-122, 2008.

[27] Fecher, N. and Watt, D., "Consonant confusions in the perception of speech through facewear", presented at the BAAP 2012 Colloquium, Leeds, United Kingdom, 2012.