TCD-TIMIT: An Audio-Visual Corpus of Continuous Speech

Naomi Harte, *Member, IEEE*, and Eoin Gillen

***Abstract*—Automatic audio-visual speech recognition currently lags behind its audio-only counterpart in terms of major progress. One of the reasons commonly cited by researchers is the scarcity of suitable research corpora. This paper details the creation of a new corpus designed for continuous audio-visual speech recognition research. TCD-TIMIT consists of high-quality audio and video footage of 62 speakers reading a total of 6913 phonetically rich sentences. Three of the speakers are professionally-trained lipspeakers, recorded to test the hypothesis that lipspeakers may have an advantage over regular speakers in automatic visual speech recognition systems. Video footage was recorded from two angles: straight on, and at 30°. The paper outlines the recording of footage, and the required post-processing to yield video and audio clips for each sentence. Audio, visual, and joint audio-visual baseline experiments are reported. Separate experiments were run on the lipspeaker and non-lipspeaker data, and the results compared. Visual and audio-visual baseline results on the non-lipspeakers were low overall. Results on the lipspeakers were found to be significantly higher. It is hoped that as a publicly available database, TCD-TIMIT will now help further state of the art in audio-visual speech recognition research.**

***Index Terms*—Audio-visual speech recognition.**

TABLE I
LIST OF ENGLISH-LANGUAGE AVSR DATABASES (SOME INFORMATION TAKEN FROM TABLES IN [7], [11], AND [12]) (SR = Speech Recognition)

Database Acronym	Speakers # (# Female)	Content e.g. isolated words	Video Resolution, FPS	Stated Purpose
AMP/CMU [13]	10 (3 F)	78 isolated words	720x480	N/A
AVletters [14]	10 (5 F)	Alphabet set	80x60, 25fps	Letter recognition
AV-TIMIT [15]	223 (106 F)	TIMIT-SX sentences	720x480, 30fps	Continuous SR
AVICAR [16]	86 (40 F)	Digits, TIMIT sentences	720x480, 30fps	SR in a car
AVOZES [4]	20 (10 F)	Digits, continuous speech	720x480, 30fps	Continuous SR
BANCA [17]	208 (104 F)	Numbers, names, addresses	720x576, 25fps	Speaker verification
CUAVE [18]	36 (17 F)	Digits	720x480, 30fps	Speaker-independent digit recognition
DAVID [19]	258 (126 F)	Digits, alphabet, syllables and phrases	560x480, 25fps	Speaker/SR
GRID [20]	36 (16 F)	Command sentences	720x576, 25fps	Small-vocab CSR
IBM LVCSR [21]	290	Continuous speech	740x480, 30fps	LVCSR
VidTIMIT [22]	43 (19 F)	TIMIT sentences	512x384, 25fps	AVCSR
Valid [23]	106	Digit strings + sentence	720x576, 25fps	Speaker/SR
TULIPS1 [24]	12 (3 F)	First 4 English digits	100x75, 30fps	Isolated digits
XM2VTS [25]	295	Digit strings + sentence	720x576, 25fps	Speaker/SR
QuLips [8]	N/A	Digit strings + sentence	720x576, 25fps	
CMU-AVPFV [26]	10	150 isolated words	640x480, 30fps	Profile vs front view lip features
HIT-AVDB-II [12]	30 (15 F)	Digits, English and Chinese phrases	N/A	View angle for speaker and SR
LiLIR [27]	1	200 sentences	N/A	View angle for SR
WAPUSK20	20 (9 F)	Command sentences	640x480, 32fps	
BAVCD [28]	15	Connected digits	640x480, 20fps	Visual and depth feature examination
UNMC-VIER [29]	123 (49 F)	digits, TIMIT sentences	708x640, 25fps	Environments and SR
AusTalk [30]	1000	Digits, isolated words, SCRIBE sentences	640x480, 48fps	Speaker/SR

I. INTRODUCTION

THE aspiration to incorporate visual information into speech recognition systems to improve robustness is well established. The first joint audio-visual speech recognition (AVSR) system was developed by Petajan in 1984 [1]. The system improved speech recognition results on a single-speaker, 100-word task. Since then, researchers have demonstrated the benefits of AVSR over audio-only speech recognition in a variety of tasks [2]. However, as observed by Potamianos *et al.* in that paper, it is difficult to compare the many algorithms that have been suggested, as they are "rarely tested on a common audio-visual database". Databases remain problematic in AVSR research today, over a decade later. The lack of suitable databases is a commonly-cited issue in AVSR research [2]–[8].

A list of commonly-used English language AVSR databases is given in Table I. Only five of these databases are suitable for medium to large-vocabulary continuous speech recognition: AV-TIMIT, GRID, VidTIMIT, IBM LVCSR and AusTalk. Of these, only GRID and VidTIMIT are currently available: AV-TIMIT and IBM LVCSR have not been released, while AusTalk is not yet available though a release is planned. Of the remaining two, GRID is larger than VidTIMIT (1000 sentences vs. 430) and filmed at a higher resolution, but its vocabulary (51 words) is much smaller as the database was primarily developed with speech intelligibility research in mind. With a limited vocabulary, the coverage of visemes is not well balanced (e.g., a digit task like CUAVE only covers 9 visemes with limited contexts). Thus, whilst a passing glance at Table I may suggest there are many datasets for continuous AVSR research, this is not borne out in practice. This paper presents work on the creation of a new audiovisual database for AVSR research.

Section II considers the desirable attributes of a database for AVSR research. Section IV gives insight into the recording and post-processing required to deliver 13826 clips in MP4 format from the 62 speakers, as well as a full set of phoneme and viseme-level transcriptions. Baseline performance is reported in Section VII with analysis of the performance of both the so-called volunteers (normal-speaking adults) and the professionally trained lipspeakers. This database is being made publicly available to help further state of the art for audio-visual speech recognition. It is envisaged that the database will be particularly valuable for research focused on improving visual feature extraction for AVSR.

Manuscript received September 26, 2014; revised February 06, 2015; accepted February 09, 2015. Date of publication February 26, 2015; date of current version April 15, 2015. This work was supported by Science Foundation Ireland under Grant 07/EN/E007. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Gokhan Tur.

The authors are with the Sigmedia Group, Department of Electronic and Electrical Engineering, Trinity College Dublin, Dublin 2, Ireland (e-mail: nharte@tcd.ie; ogiollae@tcd.ie).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TMM.2015.2407694

II. Designing an AVSR Database

What should the "ideal" AVSR dataset contain? While developing the DUTAVSC Dutch AVSR corpus, Chitu and Rothkrantz [9] undertook a review of corpora designed for AVSR. They made a number of design recommendations for any new AVSR database. Among their findings, they list common limitations found in the databases they reviewed.

- The recordings contain only a small number of respondents.
- The pool of utterances is usually very limited.
- The quality of the recordings is often very poor.
- The datasets are not publicly available.

Gan [6] and Cappelletta [7] also listed ideal features for an AVSR database. From these authors, as well as the motivations given for AVSR corpora produced in the last decade, there is good consensus between researchers on desirable features. The recurring themes in requests are:

- a large number of speakers;
- continuous speech with good coverage of phonemes and visemes;
- available to other researchers;
- high-quality recordings;
- gender-balanced speaker set.

These requests provided the catalyst for the new audio-visual continuous speech recognition database introduced in this paper. The database is named TCD-TIMIT as it was developed in Trinity College Dublin in Ireland, and its speech material is sentences from TIMIT [10], an audio-only speech recognition database created in the 1980s. The database contains audio and video footage (from two angles) of 59 volunteers, as well as 3 professional lipspeakers (see Section III). Volunteers say 98 sentences each, while the lipspeakers say 377 sentences each. Audio, visual, and audiovisual speech recognition baselines are also provided with the database. The intention is that this freely available dataset will act as a catalyst for improved visual feature extraction in AVSR research. The paper also details some of the practical issues involved in the audio-visual database construction, which will provide a useful resource for other researchers embarking on a similar task.

III. Human Lip Reading

Human speech recognition is bimodal in that humans fuse audio and visual signals to determine what a speaker says. Lip reading is a skill subconsciously used by most humans, especially when it is difficult to hear a speaker. The term "speech reading" is preferred by experts[1] as many other details about the speaker (e.g., eyes, mood, context etc) are also used by speech readers. Some deaf societies offer classes teaching people how to improve their speech reading. There are also organizations offering professional training to people to make them easier to lipread. Individuals who have completed this training are called lipspeakers, and can be used as translators for speech readers.[2] An information leaflet published by the Victoria Deaf Society lists some of the things that their lipspeakers are trained to do to make their visemes more distinctive.

1) Increase the duration of the sound /m/ to distinguish it from /p/, /b/.
2) Increase the duration of the sound /n/ to distinguish it from /t/, /d/.
3) Place the tongue between the upper and lower teeth to clarify /th/.
4) Spread the lips, clench the teeth firmly and grin to indicate /s/, /z/.
5) Bite the lower lip with the upper teeth to indicate /f/, /v/.
6) Move the jaw down briefly while producing /k/, /g/.
7) Shrug the shoulders briefly during the inhalation preceding /h/.
8) Increase or decrease height/width of the lip opening while producing vowels.

Thus rather than exaggerating mouth movements, lipspeakers attempt to make their mouth movements more distinctive. Since human lipreaders find them easier to lipread than regular speakers, it is hypothesized that they may provide insight as to the best features to use for visual speech recognition. As the purpose of the TCD-TIMIT database is to help investigate visual features for ASR, it was deemed useful to include some lipspeakers in the database along with non-trained speakers.

IV. Database Recording

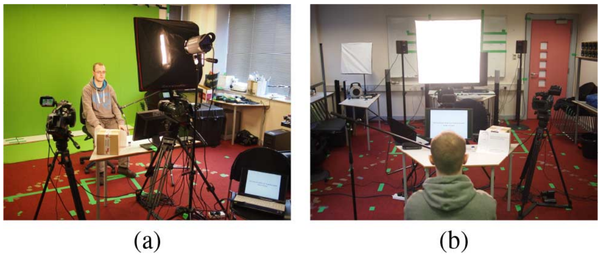
A. Speaker Scripts

The TIMIT sentences [31] were chosen as the speech material for the TCD-TIMIT corpus. Four of the databases in Table I have used groups of sentences from TIMIT as these sentences were originally designed to include as many phoneme pairs as possible. This offers correspondingly high viseme coverage. TIMIT's sentence list consists of two "SA" sentences, designed to highlight a speaker's accent, 450 "SX" sentences, hand-designed to include as many different pairs of phonemes as possible, and 1890 "SI" sentences, picked from playwrights' books to include phoneme pairs in "unusual" contexts. Speakers in TIMIT said 10 sentences each: the two SA sentences, five SX sentences and three SI sentences. The intention for TCD-TIMIT was to have speakers say much more than 10 TIMIT sentences each. This motivated an examination of the phonetic balance of groups of sentences picked from TIMIT. If a TCD-TIMIT speaker was given a particular number of sentences from TIMIT to say, would they adequately cover all phonemes? All visemes?

After trial runs with early volunteers and script sizes of 50 to 200 sentences, a script size of 98 sentences per speaker was identified as optimal. Feedback from the early volunteers suggested that this was the most sentences potential volunteers could be asked to say without discouraging them from volunteering. The rate of occurrence of all phonemes and visemes in each speaker script is consistent with the overall occurrence rates in TIMIT, giving good phoneme and viseme converage. To make scripts of 98 sentences, the TIMIT speaker scripts were split into groups of 12. Thus a single speaker spoke 8 sentences from each of the twelve TIMIT speakers in a group (i.e., 96 sentences), plus the SA1 and SA2 sentences.

[1] "Victorian deaf society information sheet," [Online]. Available: http://www.vicdeaf.com.au/files/editor_upload/File/Information%20Sheets/Speechreading%20_Visual%20Cues_.pdf

[2] "Lipspeakers," [Online]. Available: www.deafhear.ie/DeafHear/lipSpeakers.html



(a) (b)

Fig. 1. View behind the camera (a) and behind the speaker (b).

B. Equipment

A pair of Sony PMW-EX3 cameras were used to record the database. The PMW-EX3s can be synchronized, making time-alignment of the two camera views much easier after recording. They also have inputs for recording external microphones, meaning that the audio and video streams were also synchronized. The cameras were set to record 1920x1080-pixel frames at 30fps. The cameras can record at 50fps, but this is only available for 1280x720-pixel frames. A study by Saitoh and Konishi [32] found that a higher frame rate did not lead to additional improvements in visual recognition scores. As a result, higher resolution frames were chosen over a higher frame rate. For the microphone, a wireless clip-on electret mic was used. The mic was a Shure PG185, with a PG1 transmitter and a PG4 receiver. The room was not soundproof, so other mics were picking up external noise. The clip-on had some high-frequency hiss, but picked up the least external noise. The volunteers clipped the mic on themselves, with the instruction that it be below the chin, close to the mouth, and angled towards the mouth. Audio levels were then checked.

C. Recording Setup

The setup of the room is shown in Fig. 1. One camera recorded the speaker from directly in front, while the other recorded at an angle of 30° to the speaker's right. The ideal angle for viewing a speaker's face for lipreading in the context of audio-visual speech recognition is not established. A frontal view offers information about mouth width, but a profile view offers information about lip protrusion. Angles in between have also been tested in an attempt to find a "sweet spot". The debate must also take practical use cases into account. For example, a recognizer designed to be used with laptop webcams should be trained on frontal views. A recognizer to be used by car drivers, however, might have to work with angled views, as the camera cannot block the driver's view.

An investigation into the performance of profile versus frontal views was undertaken by Lucey and Potamianos [33] in 2006. Using DCT features, they found that frontal views significantly outperformed profile views on a visual-only word-recognition task. Kumar *et al.* [26] later compared profile and frontal views using shape-based features, and found that profile views gave the highest performance. In 2010, Pass *et al.* [8] investigated the viability of pose-invariant visual only speech recognition on a speaker-dependent, isolated-digit database using DCT features. Using recorded angles from $0° - 90°$, they found that the best set of DCT coefficients to retain in a frontal-view trained recognizer were angle-dependent. In 2012, Lan *et al.* [27] put together a small, single-speaker database with camera angles of 0, 30, 45, 60 and 90 degrees. Using an Active Appearance Model (AAM), they found that the view from 30 degrees gave the highest performance. Based on these results, it was decided to record speakers in the TCD-TIMIT database from two views: frontal and 30 degrees.

Both cameras were zoomed in to contain only the speaker's head, shoulders and the greenscreen in the shot. The light used was a 500W tungsten photoflood with a 60x60 cm softbox. The light was placed directly behind and above the front-facing camera and angled slightly downwards (about 15 degrees) at the subject. This position was chosen to get even lighting across the subject's face, illuminate as much of the face as possible and eliminate shadows on the greenscreen behind the subject.

D. Typical Recording Session

A typical recording session played out as follows. The speaker arrived and was given a short overview of what was involved. They were seated, mic'd and the cameras were set up. Audio levels, lighting and camera parameters were then checked. The speaker was told to try their best to start and end each sentence with their mouth closed and to leave two seconds of silence between each sentence. If the speaker made a mistake during a sentence attempt, they were directed to leave two seconds of silence before attempting it again. The only people present during the recording session were the speaker and the recorder. It was the recorder's job to advance the speaker's "teleprompter" (an external monitor hooked up to a laptop controlled by the recorder). Since the recorder could also see the sentences, it was their job to catch any pronunciation mistakes made by the speaker. After the recording process, the volunteers signed a consent form. An example of the consent form is included in the release.

V. Post-Processing

The Sony PMW-EX3s record video in MPEG-2 Long GoP format, and audio in stereo PCM, onto SxS cards, Sony-designed flash memory cards compliant with the ExpressCard standard. The video and audio are both wrapped in an MP4 container. The cameras cannot record single files larger than 4 GB, so long recordings are technically split into two or more MP4 files connected by a third file explaining the structure.

With a downloadable Sony driver, the SxS cards can be recognized through a laptop's ExpressCard slot. A minute of footage from one of the cameras is roughly 260 MB. The average footage length was roughly 15 minutes. After recording 59 speakers, 450 GB of raw footage had been recorded.

The goal of post-processing was to obtain a set of video and audio clips, one for each sentence, preferably in the exact same format as that taken from the cameras, or at the very least with no loss of quality. It was not desirable to clip the footage manually due to the time and potential for error involved. Also, while Adobe Premiere (the editor that would have been used to clip the footage manually) had no problem reading in the Sony footage and recombining the two or more MP4 files that made up a long recording session, it could not output it in the same format. Several lossless codecs were tried, but the output file sizes were too large to be practical.

Attention then turned to ffmpeg,[3] the command-line video editing tool. Since it was called from the command line, it could be automated, but unfortunately it could not process the Sony MP4 files correctly, due to an inability to wrap PCM audio into an MP4 container. However, it could wrap MPEG-2 video and PCM audio into a Matroska container (MKV), an open-source, free container format. Matroska videos can also be clipped by ffmpeg. Thus MKV became the file format used for the sentence clips during the baseline experiments. This format was changed for the release version of the database (see Section V-A).

A. Clipping the Footage

The purpose of clipping the footage was to create a clip from the straight and 30° camera for each sentence. Since the clip-on mic channel was the best-quality audio channel recorded, it was used as the audio for the 30° camera footage as well. This meant synchronizing the 30° camera footage with that of the straight camera. In addition to being able to read Sony MP4 files, Adobe Premiere can read synchronization information between them, but as explained, cannot output the footage without transcoding. The workaround solution was to synchronize the videos in Premiere, export the synchronized audio, then use ffmpeg to combine that audio and the video from the 30° camera in an MKV file. In this way, the 30° camera footage could be clipped in the same way as the straight camera. This means that each sentence clip from the straight camera has a corresponding clip from the 30° camera shot at the exact same time and with the exact same audio track.

Once the audio was synchronized and extracted, a very basic speech detector based on energy thresholding, was used to find the beginning and end of speaker sentences in the audio. Every clip was then manually checked to see if it was a valid sentence. Missing or incomplete sentences were manually re-clipped. Once all of the correct sentence clips had been identified and created, they were given their TIMIT sentence codes. A corresponding audio clip was created for each video clip (using ffmpeg) for audio-only speech recognition tests. It was later discovered that ffmpeg had not clipped the audio and video at the same places. This was due to ffmpeg being prohibited from transcoding the video and audio streams while clipping, which forced it to clip the visual stream at the nearest MPEG-2 I-frame. The audio and video appear to be in-sync when viewing the resultant clips, as the audio-visual offset is written in the meta-information of the MKV files and is taken into account by standard video players. The issue was only discovered while trying to obtain a viseme recognition baseline using viseme label files created from the time-aligned phoneme label files.

TABLE II
TIMIT PHONEME SET (OF LEE AND HON [41]) VERSUS CMU PHONEME SET

TIMIT	CMU	TIMIT	CMU
aa	AA0, AA1, AA2	ae	AE0, AE1, AE2
ax-h, ah	AH1, AH2	ax	AH0
ao	AO0, AO1, AO2	eh	EH0, EH1, EH2
ix, ih	IH0, IH1, IH2	ey	EY0, EY1, EY2
ay	AY0, AY1, AY2	iy	IY0, IY1, IY2
oy	OY0, OY1, OY2	aw	AW0, AW1, AW2
ow	OW0, OW1, OW2	ux, uw	UW0, UW1, UW2
uh	UH0, UH1, UH2	axr, er	ER0, ER1, ER2
eng, ng	NG	sh	SH
ch	CH	jh	JH
zh	ZH	y	Y
dh	DH	p	P
b	B	em, m	M
dx, t	T	d	D
en, nx, n	N	k	K
g	G	s	S
z	Z	f	F
v	V	w	W
el, l	L	r	R
th	TH	hv, hh	HH
#h, h#, epi, pau, kcl, tcl, pcl, gcl, dcl, bcl	sil	q	removed and time added to following phoneme

In conducting the visual and audio-visual baseline experiments in this paper, the audio-visual offset was found for each clip and applied to each viseme label file. Clearly, releasing the database in this state would be impractical, so the decision was taken to re-clip every sentence, this time allowing ffmpeg to transcode the video and audio using the H.264 encoder libx264. This is the final release format of the sentence clips in the database.

B. Normalizing the Audio Clips

The LUFS scale, a metric developed by the International Telecommunications Union (ITU) and known as Loudness Units relative to Full Scale (LUFS), was used to compare the loudness of the clips. The LUFS scale takes human perception of loudness, periods of silence and other factors into account. Its implementation is described in the ITU's recommendation paper BS.1770 [34]. LUFS have been adopted by the European Broadcasting Union (EBU) as a standard. Adobe Audition has an implementation of the LUFS standard which allows clips to be analyzed and normalized to a certain LUFS level. Using Audition, the average loudness of the original TIMIT clips was found to be -21.3LUFS. The TCD-TIMIT audio clips were then normalized to this level.

C. Phoneme-Level Label Files

Phoneme-level label files contain the list of phonemes in a sentence's audio clip, along with their start and end times. They are used as ground-truth for training and testing an audio-only speech recognizer. Using a phoneme-to-viseme map, they can also be converted into viseme-level label files and used to train and test a visual speech recognizer. The original TIMIT database includes word-level and phoneme-level label files for all sentences. Most published work on TIMIT uses a reduced set of phonemes (shown in Table II), and this reduced set is used in TCD-TIMIT.

TIMIT's time-aligned phoneme label files were created by hand, by phoneticians. Ideally, TCD-TIMIT's phoneme label files would be created in a similar manner, but this is painstaking, time-consuming work. To overcome this problem, researchers commonly make use of a technique called forced alignment to obtain time-aligned label files of acceptable quality ([35]–[39]). The two most widely-used toolkits for

[3] [Online]. Available: http://ffmpeg.org

TABLE III
HIBERNO-ENGLISH RULES FOR CMU DICTIONARY

Example Word	Original Phonetic Transcription	Rule Applied	Alternate Phonetic Transcription
this	DH IH1 S	DH ->D	D IH1 S
I'm	AY1 M	AY ->OY	OY1 M
three	TH R IY1	TH ->T	T R IY1
validate	V AE1 L AH0 D EY0 T	T ->AH1	V AE1 L AH0 D EY0 AH1
butter	B AH1 T ER0	AH1 AH1 ->UH0	B UH0 ER0

automatic speech recognition, HTK and CMU Sphinx, provide forced alignment capabilities.

The idea behind forced alignment is that if there are well-trained phoneme HMMs available, they can be exploited to provide the maximum likelihood of producing some non-aligned observation sequence. The phoneme boundaries can then be taken as the boundaries between the HMMs. The key here is that the phonemes in the sequence, and their order, are known. Hence the HMMs to use, and order in which to arrange them, is known. The only missing information is the boundaries between the phonemes.

The "Penn Phonetics Lab Forced Aligner" (P2FA) is a tool developed by Yuan and Liberman [40] during their research into identifying speakers in the SCOTUS (Supreme Court of the United States) corpus. The corpus contains over 9000 hours of oral arguments from the Supreme Courts. With 25.5 hours of manually-aligned training data, they built a recognizer capable of force-aligning long (1 hr+) segments of speech. Testing on a manually-aligned subset, they found that the majority of the differences between the manually and force-aligned boundaries were under 50 ms. With this in mind, it was decided to obtain force-aligned files for TCD-TIMIT using P2FA.

The P2FA tool consists of a set of HTK-compatible models and other files, and a Python script which sets up and calls HTK in forced-alignment mode. The Python script produces a force-aligned phoneme-level file for a given speech clip and its word-level label file (it disregards time information). It does this by using a pronouncing dictionary, specifically the CMU pronouncing dictionary,[4] which contains transcriptions for over 125000 words. The pronouncing dictionary is used to get the transcriptions for the words in the word-level file. The phonemes are then force-aligned. The advantage of using word-level transcriptions and a pronouncing dictionary over force-aligning the TIMIT phoneme-level transcriptions is that the pronouncing dictionary can have multiple transcriptions for the same word. HTK tries all of these and picks the most accurate one. This means that different pronunciations can be accounted for by having multiple transcriptions for words.

The CMU dictionary had multiple pronunciations for some words, but they were all based on American English. There was no allowance for Hiberno-English [42], particularly Dublin accents. Based on the traits of Hiberno-English, a set of rules was created and applied to the CMU dictionary to add Hiberno-English transcriptions of words that fell under one or more of the rules. A summary of these rules is shown in Table III. After this process, over 29000 extra transcriptions had been added to the CMU dictionary.

In addition, a very small number of words in the TIMIT sentences were not present in the CMU dictionary. Since there were only a few of them, the missing words were manually given CMU-compatible transcriptions, following the mapping in Table II. In cases where stress indicators were needed, similar words were consulted for a suitable indicator to apply. The newly-transcribed words were inserted into the CMU dictionary.

P2FA's Python script needed minor modifications to run in Windows, as it was developed for Linux machines. Calls to Unix utilities cp, rm and cat were changed to their Windows equivalents. Calls to the SoX sound processing program were replaced with calls to ffmpeg. Since it only processes one sound clip at a time, a second wrapper script was written to run it for each TCD-TIMIT clip. Each transcription was added to one large text file.

After the transcriptions had been obtained, they were mapped back to TIMIT-compatible phonemes using the reduced TIMIT phoneme set. This allows the use of the phoneme-to-viseme map based on the reduced TIMIT phoneme set. The map of Table II was used for this purpose. The only TIMIT phoneme that did not have an equivalent in the CMU dictionary was /dx/, so this phoneme was ignored during the reverse mapping and not used afterwards. The only other issue was the stress indicators in the CMU vowels, which were simply discarded. After the reverse mapping, the file containing the transcriptions was turned into a HTK-compatible MLF (Master Label File).

D. *Viseme Labels*

Dominant approaches to automatic visual speech recognition rely on an analogue to phonemes called visemes. Unfortunately visemes are not as well defined as phonemes. There are two competing definitions, with neither clearly superior.

The first definition (a data-driven approach) is based on articulatory gestures: Visemes can be thought of in terms of articulatory gestures, such as the lips closing or rounding, teeth exposure, jaw movement etc., without a link to the uttered phoneme. Hence there is no defined link between the phonemes and visemes.

The second definition (a linguistic approach) is based on the corresponding phonemes: Visemes are derived from groups of phonemes having the same visual appearance. This definition allows for phoneme-to-viseme maps, which are many-to-one maps, as some phonemes are visually indistinguishable.

Note that neither definition suggests that visemes distinguish words, as phonemes do. In reality, a viseme is somewhere in-between. The natural asynchrony of the visual and audio signal in uttered speech still allows for a close relationship between phonemes and the visemes and it is likely that humans exploit these cues. The phoneme-to-viseme mapping allows for a convenient extension of standard ASR frameworks based on Hidden Markov Models and has been well reported in the literature. For the reported baselines with this database release, a phoneme-to-viseme mapping based on a map designed by Jeffers and Barley [43] is used. The map is given for TIMIT's phoneme set in Table IV.

Two TIMIT phonemes are not present in the map: /hh/ and /hv/. These two phonemes do not have any viseme associated with them, as a speaker produces these phonemes while

[4] [Online]. Available: http://www.speech.cs.cmu.edu/cgi-bin/cmudict

TABLE IV
JEFFERS AND BARLEY MAP, INCLUDING VISIBILITY AND OCCURRENCE RATES

Viseme	TIMIT Phonemes	Description	Visibility Rank	Occurrence [%]
/A	/f/ /v/	Lip to Teeth	1	3.17
/B	/er/ /ow/ /r/ /q/ /w/ /uh/ /uw/ /axr/ /ux/	Lips Puckered	2	15.49
/C	/b/ /p/ /m/ /em/	Lip Together	3	5.88
/D	/aw/	Lips Relaxed-Moderate Opening to Lips Puckered-Narrow	4	0.7
/E	/dh/ /th/	Tongue Between Teeth	5	2.9
/F	/ch/ /jh/ /sh/ /zh/	Lips Forward	6	1.2
/G	/oy/ /ao/	Lips Rounded	7	1.81
/H	/s/ /z/	Teeth Approximated	8	4.36
/I	/aa/ /ae/ /ah/ /ay/ /ey/ /ih/ /iy/ /y/ /eh/ /ax-h/ /ax/ /ix/	Lips Relaxed Narrow Opening	9	31.46
/J	/d/ /l/ /n/ /t/ /el/ /nx/ /en/ /dx/	Tongue Up or Down	10	21.1
/K	/g/ /k/ /ng/ /eng/	Tongue Back	11	4.84
/S	/sil/ /pcl/ /tcl/ /kcl/ /bcl/ /dcl/ /gcl/ /h#/ /#h/ /pau/ /epi/	Silence	-	-

forming the next viseme with their mouth. As a result, when these phonemes appear in a transcription they are mapped to the next viseme in the sequence.

E. Region of Interest Extraction

A set of extracted "region of interest" frames or ROIs (see [2]) is provided with the database. Researchers may use these or experiment with alternative ROIs for visual feature extraction. The set provided were extracted using a nostril tracking algorithm developed by Cappelletta [7]. The algorithm, originally developed for VidTimit, worked on about 85% of the TCD-TIMIT footage. Unfortunately, unlike Cappelletta's reported findings, mouth detection did not work on all of the clips where nostril tracking was successful. The vast majority of the nostril and mouth failures were due to similar problems and contained within a small group of speakers: 13F, 25M, 26M, 29M, 30F, 34M, 40F, 52M, 56M and 57M. Small modifications to the original algorithm, including restraining the searched nostril region, resolved the issues. Seven other clips were manually tracked. In each of the 7 clips, the speaker had moved their head rapidly at some point and the algorithm lost track of the nostril region. These speakers and their associated clips will be of interest to researchers looking at fully automated ROI extraction or improved tracking. The visual and audio-visual baselines reported in this paper use the ROIs supplied with the database release.

VI. FINAL DATABASE STRUCTURE

In total, there are 62 speakers in TCD-TIMIT. Three of these are professional lipspeakers. All the lipspeakers are female. The average age of the lipspeakers is 60, with a variance of 2. The other 59 TCD-TIMIT speakers are non-lipspeakers and were recruited from volunteers around the University. Of these, 32 are male and 27 are female. The average age is 24, with the minimum age 16 and the maximum age 57. All speakers signed consent forms to allow for research use of the database.

TABLE V
RECOMMENDED TCD-TIMIT TRAIN-TEST SPLIT

TRAIN	24M	04M	26M	02M	32F	47M	06M	50F	59F	23M
	19M	05F	31F	22M	01M	39M	46F	11F	42M	57M
	43F	29M	17F	37F	21M	12M	38F	48M	16M	52M
	40F	13F	14M	03F	20M	51F	30F	10M	07F	
TEST	28M	55F	25M	56M	49F	44F	33F	09F	18M	54M
	45F	36F	34M	15F	58F	08F	41M			

TABLE VI
MONOPHONE RECOGNITION RESULTS FOR TCD-TIMIT—SPEAKER INDEPENDENT

	train set	test set
%correct	72.53	65.47
%accuracy	57.82	47.63

TABLE VII
MONOPHONE RECOGNITION RESULTS FOR TCD-TIMIT—SPEAKER DEPENDENT

	train set	test set
%correct	72.32	66.81
%accuracy	57.23	49.84

The speaker's identification code format was chosen to make it as easy as possible to iterate through speakers. Speakers have a two-digit ID code, from 01-99 with the letter "M" or "F" appended to indicate whether a speaker is male or female. The reason TIMIT's speaker codes were not re-used, as they are in VidTIMIT, is that TCD-TIMIT speakers say more than one TIMIT speaker's sentences, so there is no direct relationship. Speaker 27M has a Spanish accent, while 35M and 53M have British accents. The rest of the accents in the database are Irish accents, the majority being "neutral" Dublin accents. The data was shot in front of a green screen for possible speaker segmentation applications as in CUAVE. Speakers were allowed to wear glasses, piercings and any hairstyle, the only request was that nothing green be near the head or shoulders. This was to make colour-based segmentation easier. Speakers were instructed to speak in the most natural way possible, i.e., they had control over speed, pauses, intonation etc. The were asked to fully close their mouth between sentences. The lipspeakers were instructed to speak as they do when providing lipspeaking services. The only difference between the recording sessions for the non-lipspeakers and lipspeakers was the length. The 3 lipspeakers said almost 4 times as many sentences as the non-lipspeakers. The complete database is downloadable from www.mee.tcd.ie/~sigmedia/Resources.

VII. TCD-TIMIT BASELINE PERFORMANCE

A. Audio-Only Baseline

A monophone recognizer was trained and tested using the audio portion of the database. Speakers 27M, 35M and 53M were not included due to their non-Irish accents. The audio files used were mono WAV files sampled at 16 kHz. The features were MFCCs with 12 coefficients extracted over a 25 ms window with 50% overlap, appended with 1st and 2nd derivatives. A 70-30 split was used in dividing the database into train and test material. The speakers in the train and test subsets are given in Table V. 3-state left-to-right HMMs were used. The number of mixtures per state was incrementally increased to 31. No language model was employed. The results are given in Table VI. For these and all subsequent tests, the terms "correct" and "accuracy" are based on the HResults tool from HTK [44]. Further tests (not reported here) verified the consistency

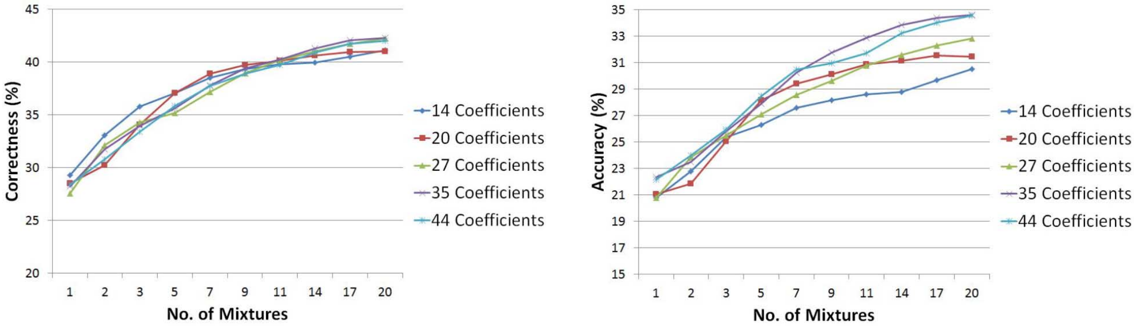


Fig. 2. Monoviseme recognition results using different combinations of DCT vector and a 4-state HMM.

TABLE VIII
VISUAL-ONLY RECOGNITION RESULTS FROM TWO SPEAKER-DEPENDENT TCD-TIMIT TRAIN-TEST SPLITS

	Split 1		Split 2	
	Train set	Test set	Train set	Test set
%correct (%)	42.69	41.98	42.19	41.89
%accuracy (%)	36.05	34.54	36.01	34.82

TABLE IX
TEN BEST AND WORST SPEAKERS FROM THE 4-STATE, 44-COEFFICIENT EXPERIMENT OF FIG. 2

Top 10 Speakers			Bottom 10 Speakers		
Speaker	Correct	Acc	Speaker	Correct	Acc
12M:	51.9	39.67	02M:	16.68	16.49
44F:	45.21	39.71	57M:	24.92	22.89
39M:	49.56	39.86	47M:	23.83	23.24
03F:	50.87	40.1	42M:	28.6	25.9
20M:	44.88	40.24	54M:	32.97	26.86
18M:	45.96	40.37	46F:	33.3	28.69
22M:	50.44	40.39	29M:	29.61	28.77
33F:	46.02	40.99	52M:	33.39	28.78
43F:	47.1	41.1	34M:	30.8	29.22
49F:	49.62	41.9	45F:	35.85	30.14

in performance across alternative train-test divisions. This particular split is the recommended train-test division to allow for comparison of results from different researchers.

A speaker dependent audio baseline is given in Table VII for comparison. 3 state HMMs were again used with 31 mixtures for direct comparison. 70% of data from each speaker was used for training, with 30% from each speaker for testing. No adaptation was used. Performance is lower in both cases on the test set than on the training set, but not at a level that suggests over-training. The variation in performance across individual speakers is small, with many vowel confusions being attributable to vowel similarity with a Hiberno-English accent, e.g., /aw/ and /eh/. File lists for both baselines are included in the database release.

B. *Speaker-Dependent Visual Baseline*

The Discrete Cosine Transform was chosen as the feature for the baseline visual results. The DCT has been widely used in the literature and is often chosen for comparative performance for new visual feature extraction schemes. A speaker-dependent viseme recognizer was trained for every combination of 3, 4, and 5-state HMMs and 14, 20, 27, 35 and 44-length DCT coefficient vectors extracted from the supplied database ROIs. The DCT vectors were upsampled from 30 to 60fps and then concatenated with their 1st and 2nd derivatives, leading to final vector lengths of 42, 60, 81, 105 and 132. Recognition results for each vector length and a 4-state HMM are given in Fig. 2. The results for 3 and 5 states are not shown here, but both were similar with slightly lower performance than the 4-state HMM.

The results in Fig. 2 are low overall, with correctness and accuracy peaking at 42% and 34% respectively at the highest mixture count of 20. The performance of the different vector lengths is very similar as the number of mixtures and states are increased. However, the 14 and 20-coefficient vector scores are consistently below the others at 20 mixtures. This is consistent with results found by others. Heckmann *et al.* [45] found no decrease in word error rate from varying the number of DCT coefficients between 20 and 100 on a single-speaker isolated digit recognition task. Seymour *et al.* [46] tested two different methods of extracting the most relevant DCT features on a speaker-independent isolated digit recognition task. They found the performance of both methods to be similar beyond 40 coefficients. Scanlon *et al.* [47] found that of vectors containing 15, 28 and 36 DCT coefficients and their deltas, the 28-coefficient vectors gave the highest performance on an isolated-word recognition task. Their deltas were not calculated between adjacent frames but between frames a certain distance apart (distance depending on the length of the utterance). They also found that using 28 deltas alone led to higher word-recognition accuracy than any combination of DCT coefficients and deltas. They note that “increasing the number of transform coefficients increases the visual recognition accuracy” but also note that “the size of training data available limits the possible feature vector dimensions for good recognition”.

The findings of these researchers and the results of Fig. 2 suggest that the 27, 35 and 44-length DCT coefficient vectors are all adequate for representing the ROIs. Since the 44-length vector has consistently high accuracy, it was chosen as the vector length for further visual-only and audio-visual experiments. All subsequent experiments in this section use 4-state HMMs and 44-coefficient DCT vectors (plus 1st and 2nd derivatives) unless otherwise specified.

Overall, the performance demonstrated in Fig. 2 is low. Performance at the level reported by Heckmann *et al.* [45] or Seymour *et al.* [46] was not expected, since isolated digit recognition is a simpler task than viseme recognition in continuous speech. Cappelletta [7] reports on a speaker-dependent viseme

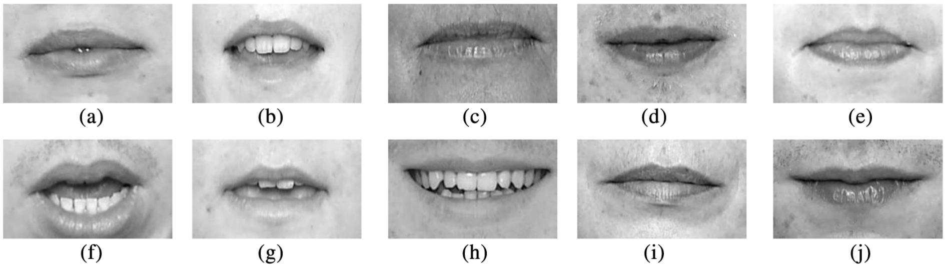


Fig. 3. Sample ROIs from the top ten speakers of Table IX. (a) 49F. (b) 43F. (c) 33F. (d) 22M. (e) 18M. (f) 20M. (g) 03F. (h) 39M. (i) 44F. (j) 12M.

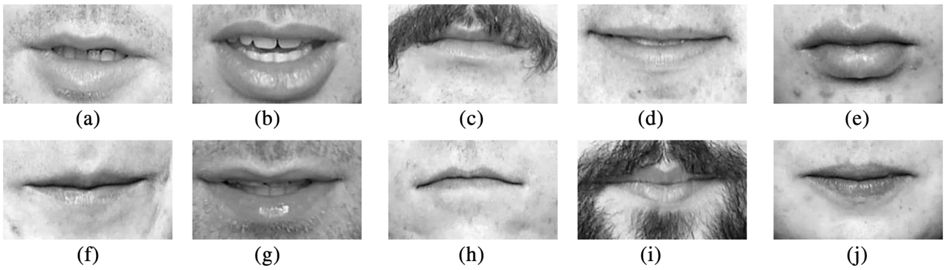


Fig. 4. Sample ROIs from the bottom ten speakers of Table IX. (a) 02M. (b) 57M. (c) 47M. (d) 42M. (e) 54M. (f) 46F. (g) 29M. (h) 52M. (i) 34M. (j) 45F.

TABLE X
VISUAL-ONLY VISEME RECOGNITION RESULTS FOR VECTORS COMPOSED OF DIFFERENT DCT COMPONENTS

Components	Length	Corr	Acc
Original Coefficients	44	20.76	19.69
1st Derivatives	44	42.95	28.82
2nd Derivatives	44	43.75	28
1st and 2nd Derivatives	88	48.32	30.78
Original Coefficients, 1st and 2nd Derivatives	132	41.98	34.54

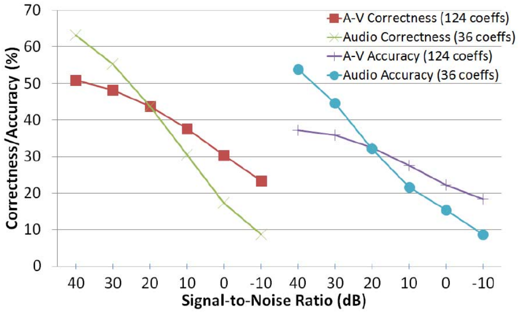


Fig. 5. Audio-visual versus audio-only, volunteers in noisy audio.

recognition experiment on VidTIMIT. He used a very similar setup (27 DCT coefficients, 4-state HMMs), and reported results of 44.89% correctness and 42.25% accuracy. The equivalent results on TCD-TIMIT are 42.19% correctness and 32.81% accuracy.

For completeness, the 4-state, 44-coefficient experiment was performed again using a different train-test split. Results for both splits are given in Table VIII. The results confirm that performance does not vary with different training and test data. The individual speaker results were then checked for significantly below-average performances. The inter-speaker performance variances were found to be high, with a correctness variance of 54.08 and an accuracy variance of 25.67. The scores of the 10 best and worst speakers are given in Table IX, and example ROIs from these speakers are given in Figs. 3 and 4 for comparison. The images of Figs. 3 and 4 do not provide any especially obvious clues to explain the performance variance between the best and worst speakers overall. It is interesting to note though that the two volunteers with the most facial hair are in the bottom 10. Also worth noting is that the top 10 speakers contain an equal number of males and females, while the bottom 10 consists of 8 males and 2 females. The average accuracy of the 29 male speakers was 32.93%, while the average accuracy of the 27 females was 36.21%. One theory behind the disparity between the top and bottom 10 is that most of the top 10 speakers were quite expressive, whereas most of the bottom 10 appear to move their mouths less while speaking. This is only a subjective judgment though and requires more investigation. The full scores of every speaker are given with the database release.

C. *Audio-Visual Baseline*

For TCD-TIMIT's audio-visual baseline, early integration was chosen. This involved concatenating the MFCC audio feature vectors with their DCT visual feature vector counterparts. To concatenate audio and video feature vectors, their framerates must be equal. One way to accomplish this is to upsample the video feature vectors to the audio framerate using linear interpolation ([6], [48]–[50]). This method was used to upsample the DCT vectors to 100FPS, the audio sample rate. Since 44-coefficient DCT vectors had been found to perform well in Section VII-B, these were chosen as the DCT vectors to concatenate with the MFCCs. However, the full length of these DCT vectors was 132, since they also contained the 1st and 2nd derivatives of the coefficients. Concatenating these vectors with the MFCCs would have produced 168-length vectors, considered too large for the amount of training data available. To reduce the vector length, tests were undertaken to find the most useful components of the DCT vectors. Visual-only recognizers were trained and tested on several combinations of

TABLE XI
DIFFERENCES BETWEEN AUDIO-VISUAL (AV)/AUDIO-ONLY (A) ACCURACY AVERAGES OF VISUAL-ONLY BEST/WORST TEN (FOUND IN TABLE IX) AND OVERALL AVERAGE AT EACH SNR. POSITIVE DIFFERENCES ARE ABOVE-AVERAGE

SNR	40dB		30dB		20dB		10dB		0dB		-10dB	
	AV	A	AV	A	AV	A	AV	A	AV	A	AV	A
Average %	37.24	53.88	35.95	44.63	32.41	32.2	27.55	21.6	22.3	15.49	18.33	8.73
Diff. from av % (Best 10)	1.9	0.52	1.94	0.02	2.29	0.23	1.96	0.4	1.29	-0.09	1.5	0.1
Diff. from av. % (Worst 10)	-2.32	-0.89	-1.4	2.32	-1.21	2.41	-1.62	0.89	-1.43	0.71	-1.82	0.06

original coefficients, 1st and 2nd derivatives, and their results were compared. These results are given in Table X. From the table, the closest performance to the original 132-length vectors was obtained on vectors consisting of the 1st and 2nd derivatives of the DCT coefficients. Hence, these were chosen as the visual component to concatenate with the audio MFCCs, leading to a final audio-visual vector length of 124.

D. Audio-Visual Speech Recognition in Noise

The most common experiment reported on audio-visual speech recognizers is to evaluate their performance as the signal-to-noise ratio (SNR) in the audio component of the signal is lowered [51], [52], [49], [47], [46]. The original test audio was corrupted by an increasing amount of additive white Gaussian noise (AWGN). A speaker-dependent recognizer was trained on the clean audio-visual data and then tested on the increasingly noisy data. For comparison, an audio-only recognizer was trained on clean audio and then tested on the increasingly noisy audio data. The results are graphed in Fig. 5.

Fig. 5 shows that after the SNR drops below 20 dB, the audio-visual recognizer's performance is higher than the audio-only recognizer. At the lowest SNR of -10 dB, the audio-visual recognizer's accuracy score is 18.32%, 10 percentage points higher than the audio-only score of 8.71%. Above 20 dB however, the audio-only recognizer outperforms the audio-visual recognizer. At 40 dB SNR (considered "clean", i.e., no AWGN was added to this audio), the audio-only recognizer's accuracy is 53.88%, almost 17 percentage points higher than the audio-visual accuracy of 37.26%. This is most likely due to the large visual component (88 of the 124 coefficients) in the audio-visual vectors. The audio-visual accuracy at 40 dB is close to the visual-only accuracy scores reported.

The trends in these results are mostly consistent with other noisy audio experiments in the literature. The most similar experiment for comparison is one by Galatas *et al.* [51]. Their audio-visual vectors also consisted of DCT coefficients (plus 1st and 2nd derivatives) concatenated with MFCC coefficients (plus 1st and 2nd derivatives), but their HMMs modelled triphones and had 4 mixtures each. Their database was CUAVE [18], hence the task was speaker-dependent isolated-word recognition on a vocabulary of 10 words. Testing at 4 SNRs (10, 5, 0 and -5 dB), they found that the word accuracy of the audio-visual speech recognizer was consistently higher than that of the audio-only recognizer. The absolute improvements at each SNR were 0.67% at 10 dB, 15.97% at 5 dB, 16.33% at 0 dB and 5% at -5 dB. This is consistent with the trend in Fig. 5, which shows the audio-visual recognizer accuracy above the audio-only accuracy at those SNRs, and shows the accuracy of both recognizers falling as the SNR decreases.

Papandreau *et al.* [52] also performed noisy audio experiments on CUAVE. Instead of MFCCs, they used 26 FBANK coefficients, and instead of the DCT, they used an Active Appearance Model. Their audio-visual integration approach was intermediate integration [2]. Their 10 HMMs modelled each word in CUAVE's vocabulary, and had 8 states each. Their audio vectors are enhanced to provide additional robustness against noise. Despite these differences, a similar trend is observed in their results, which show the audio-visual accuracy higher than the audio-only accuracy by about 15% (absolute) from -5 to 10 dB, where the gap narrows until they are within 3% of each other at the "clean" ($>$ 20 dB) SNR. Also visible is the decrease in the accuracy of both recognizers as the SNR decreases.

Finally, Zhang *et al.* [49] report correctness results using noisy audio on the AMP/CMU database [13]. For visual features they used shape-based features, while their audio features were 12 MFCC coefficients plus their 1st derivatives. They tested early, intermediate and late integration. The early integration result is the most relevant. Unlike Fig. 5, their audio-only correctness score remains below the audio-visual score over the entire range of SNRs from 30 to 0 dB. However, the gap does narrow from 45 percentage points at 5 dB to 10 percentage points at 30 dB. The results show that the performance of the audio-only and audio-visual recognizers decreases with the SNR. Direct comparison with other continuous AV speech tasks is difficult due to mismatch of experimental set-ups. Hazen *et al.* in [15] quote phone error rates of 35.2% on their AV test set in quiet but they use context, whereas this paper has only reported monophone results with no language model. Potamianos *et al.* [53] report a word-error-rate of 16% for AV concatenation compared to 14% for audio-only performance on a test set of 1038 utterances from 26 speakers in the ViaVoice scripts.

E. Individual Speaker Performances

One of the trends looked for in the individual audio-visual speaker performances was whether the best speakers from the visual-only experiments (Table IX) had higher-than-average audio-visual scores as the noise increased, and vice-versa with the worst speakers. To evaluate this, the average accuracy of the 10 best and worst speakers from Table IX is compared to the overall accuracy at each SNR in Table XI. Their audio-only performances are also given for comparison.

Table XI supports the theory that the best visual-only speakers have above-average performance in noisy conditions,

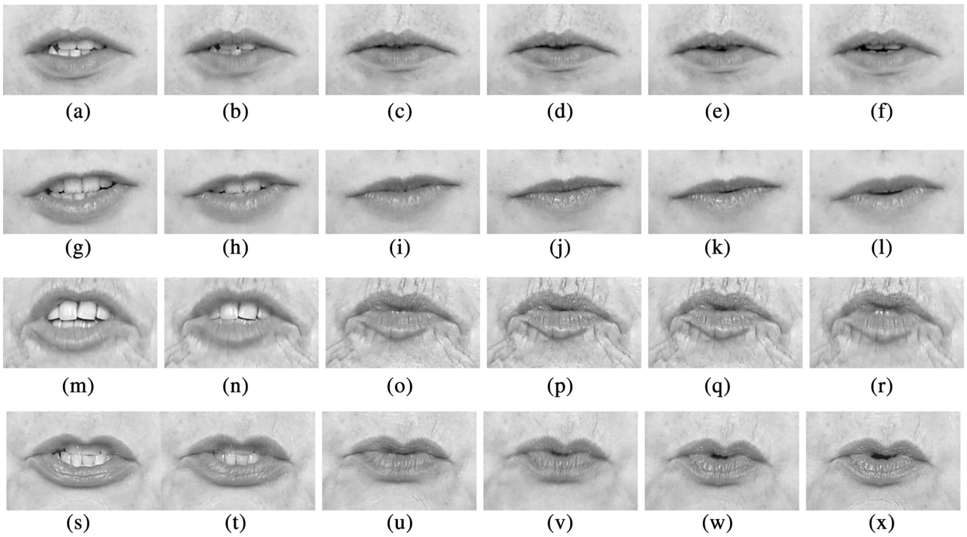


Fig. 6. ROIs extracted from two volunteers (06M and 37F) and Lipspeakers 2 and 3 for the phoneme sequence: /sh/ /w/ /aa/. This sequence occurs in the middle of the words "wash water" in TIMIT's SA1 sentence, which was said by all speakers in TCD-TIMIT. The first and second rows are volunteers (06M and 37F), the third row is Lipspeaker 2, and the fourth row is Lipspeaker 3. Note that Lipspeaker 2 is still on the /w/ phoneme by the sixth frame. The images show how the lipspeakers articulate visemes more than the volunteers. (a) /sh/. (b) /sh/. (c) /w/. (d) /w/. (e) /w/. (f) /aa/. (g) /sh/. (h) /sh/. (i) /w/. (j) /w/. (k) /w/. (l) /aa/. (m) /sh/. (n) /sh/. (o) /w/. (p) /w/. (q) /w/. (r) /w/. (s) /sh/. (t) /sh/. (u) /w/. (v) /w/. (w) /w/. (x) /aa/.

and that the converse is also true for the worst visual-only speakers. The table shows that the average audio-visual accuracy of the 10 best speakers from Table IX is higher than the overall average at every SNR, while the opposite is true for the worst 10. Looking at their audio-only scores, such differences are not apparent. The audio-only accuracy of the worst 10 is actually higher than that of the best 10 from 30 to 0 dB. Hence the better/worse audio-visual performances of the groups are due to their visual performances. A full list of individual speaker scores is given in the database release.

VIII. Lipspeaker Data

As explained in Section III, lipspeakers undergo training to become easier for human lipreaders to lipread. 3 lipspeakers were recorded as an additional part of the TCD-TIMIT database. A comparison of lipspeaker versus regular volunteer visemes can be seen in Fig. 6. The recording setup used was the same as the one described in Section IV, but the lipspeakers recorded almost 4 times as much material as the 59 volunteers (377 sentences vs. 98). Parts of the volunteers' scripts were re-used to create the lipspeaker scripts. In creating the lipspeaker scripts, the only sentences duplicated were SA1 and SA2. All of the other sentences are unique to each lipspeaker. The post-processing, creation of label files and feature extraction of the lipspeaker data was done following the same methods as the volunteer data. All experiments reported here use 4-state HMMs and 44-coefficient DCT vectors (plus 1st and 2nd derivatives) unless otherwise specified.

A. Lipspeaker Visual Baseline

A new speaker-dependent recognizer was trained and tested on lipspeaker data only. A 67-33 train-test split was used, meaning that each of the 3 lipspeakers contributed 251 sentences to the training set and 126 to the test set. The results are directly comparable to the speaker-dependent volunteer results of Table VIII.

TABLE XII
Lipspeaker Results on Lipspeaker-Only Trained HMMs Versus Volunteer Results on Volunteer-Only Trained HMMs

	Lipspeakers		Volunteers	
	Train set	Test set	Train set	Test set
%correct	60.38	57.85	42.69	41.98
%accuracy	56.45	52.74	36.05	34.54

The results of Table XII show the lipspeaker recognizer outperforming the volunteer recognizer by a large margin. Test-set accuracy is 18 percentage points higher, while training-set accuracy is 20 points higher. There are some obvious caveats to these results: the lipspeaker recognizer was trained and tested on 3 people versus the volunteer recognizer's 56, and saw far more data per speaker (251 vs. 67 sentences each). Ideally, the influence of these differences could be estimated by training a recognizer using the same amount of data from 3 volunteers, but unfortunately this data does not exist. Nevertheless, the size of the margin between the two sets of results supports the theory that lipspeakers are easier for automatic systems to lipread than volunteers. To determine whether the lipspeakers held this advantage regardless of the number of HMM states or DCT coefficients, a recognizer was trained on lipspeaker data for the same combinations of state lengths and coefficients used in Section VII-B. Graphs of the results are given in Fig. 7 for the 4-state HMM.

As was the case with Fig. 2, Fig. 7 shows that there is very little difference between DCT coefficient lengths above 20. The results for 3 and 5 states were very similar and are not shown here. The lipspeaker scores are higher than the volunteer scores in every case. This indicates that their advantage is persistent over these parameter choices.

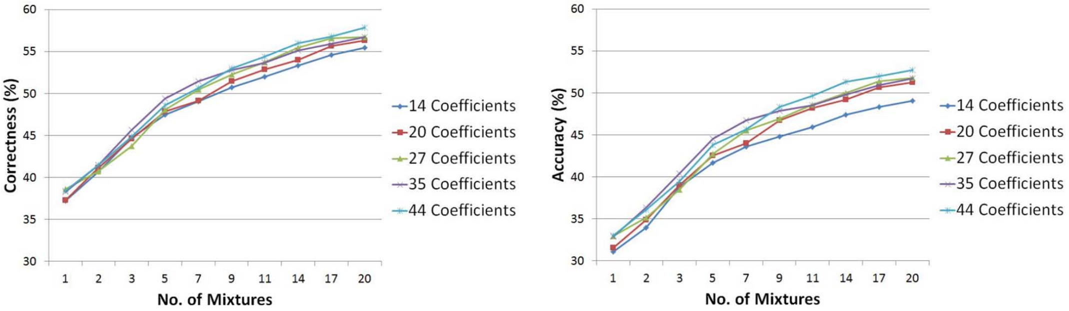


Fig. 7. Lipspeaker recognition results using different combinations of DCT vector and a 4-state HMM state.

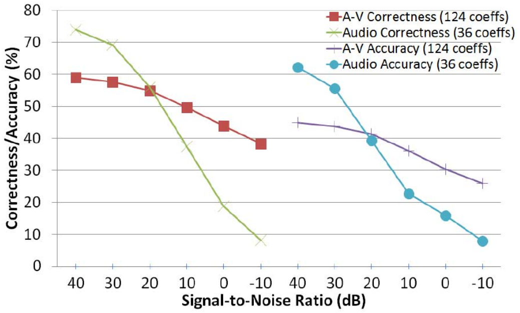


Fig. 8. Audio-visual versus audio-only, Lipspeakers on noisy audio.

There is a considerable gap between the worst and best lipspeaker scores. Lipspeaker 1 has the lowest accuracy in all three tests. One noticeable difference between Lipspeaker 1 and the others is that she spoke more slowly. Her average clip length is 6.23s, compared to 5.07s for Lipspeaker 1 and 5.13s for Lipspeaker 3. Also, the other two lipspeakers agreed that Lipspeaker 1 places the most emphasis on each viseme.

IX. Lipspeaker Audio-Visual Baseline

The final experiments run on the lipspeaker data compared audio-visual versus audio-only performance in noise. This was done on volunteer data in Section VII-D. The lipspeaker results are also compared to results from that section. The experiment settings from Section VII-D were replicated. The training set consisted of 251 sentences from each lipspeaker, leaving 126 sentences from each as the test set. The results are graphed in Fig. 8.

As Fig. 8 shows, the audio-visual recognizer begins to outperform its audio-only counterpart at an SNR of roughly 22 dB. The performance gap then widens as the SNR is lowered further. At the lowest SNR of -10 dB, the accuracy of the audio-visual recognizer is 25.86% versus the audio-only accuracy of 8.01%. At the other end of the graph, the audio-visual recognizer's accuracy with clean audio is 44.94% versus the audio-only accuracy of 62.16%.

The trends in these results are the same as those found on the volunteer data in Fig. 5. Compared to the volunteer results, the lipspeakers' audio-visual accuracy surpasses the audio-only accuracy at a slightly higher SNR (22 dB vs. 20 dB). At every SNR, the audio-visual correctness and accuracy are roughly 8% higher for the lipspeakers than for volunteers. This is understandable, considering the results of Section VIII-A. The audio-only scores, on the other hand, are roughly 8 dB higher for lipspeakers than for volunteers at 40, 30 and 20 dB. At 10 dB, the audio-only accuracy of the lipspeakers drops sharply, leaving it at a similar level (22.69%) to the corresponding volunteer score (21.58%). This occurs at 0 dB for the correctness score. As a result, at the final SNR of -10 dB, the audio-only scores for lipspeakers and volunteers are very similar at about 8% in both cases. This indicates that, in terms of audio, the benefit of fewer speakers and more training data from each speaker has been wiped out by noise at -10 dB. This further emphasizes the importance of the corresponding audio-visual results for the lipspeakers.

Confusion matrices were studied to compare trends for the volunteer and lipspeaker cohorts. These are available in the database release but not reproduced here due to space constraints. Similar-sounding phonemes with different visemes (e.g., /m/ and /n/) are confused less when visual information is introduced, while the opposite effect is observed on different-sounding phonemes with the same viseme (e.g., /g/ and /k/). Also apparent is the increased confusion between vowels and consonants which typically follow them (e.g., /ah/ and /t/, /n/ and /l/), most likely due to coarticulation effects. Comparing the volunteer and lipspeaker audio-visual results, most phonemes have higher correctness scores for the lipspeakers. The only phonemes to have higher scores for the volunteers are the least-frequent phonemes /jh/, /ch/, /ng/, /uh/, /oy/ and /th/. These phonemes were the lowest-scoring for volunteers, and the cause for this was thought to be their relative lack of training data. This theory is further supported by their even worse performance for the lipspeakers, since the amount of lipspeaker training data was lower still (753 sentences vs. 3752).

As the SNR is lowered, similar trends are seen in the lipspeaker and volunteer audio-visual results. At -10 dB, the most robust phonemes in both cases are the ones in the smallest viseme groups. As such, (/aw/, /f/, /v/, /b/, /p/) are also among the highest-performing phonemes. Deletions are once again the main factor in the audio-only results, accounting for 90% of all phonemes in the test set. Only 21% of phonemes were deleted by the audio-visual recognizer. This is even better than the volunteers' figure of 38%, and is a large factor in the improved overall score of 25.86% versus 18.32% accuracy. The high scores of some phonemes at -10 dB are inflated by large levels of insertions, particularly /hh/ and /dh/. This is understandable, since the audio profile of these phonemes is the closest to noise. However, some phonemes, for example /w/, /ae/ and /aa/, have relatively high levels of correctness without a large number of insertions, an audio profile close to noise or a small viseme group. It is probable that the lipspeakers have made these phonemes more visually distinctive than the volunteers.

X. Future Work

The TCD-TIMIT database is free for research use and down-loadable from www.mee.tcd.ie/~sigmedia/Resources. This paper introduces the database and provides useful baselines for other researchers exploring audio-visual speech. There is extensive scope for further research using the data. The primary focus in developing this dataset was for audio visual speech recognition, particularly to allow researchers improve visual feature extraction for continuous speech. At 62 speakers, the database is not as large as AV-TIMIT or IBM's ViaVoice but there is sufficient high-quality data to significantly progress research on continuous AVSR. Given the information on individuals that proved challenging for ROI extraction, the database is valuable for researchers looking at robust ROI extraction or mouth tracking. The results reported here have not used the second camera footage at 30°. Results reported for the audio visual baseline only use early integration, and assume a single direct map from phoneme to viseme. The database provides an ideal dataset to more fully explore audiovisual synchrony. Speakers were not constrained in their speaking style, thus there is a natural variation in speaker rates which has been recently focused on in [54]. The supplied labels are from a forced alignment using P2FA and have not been fully verified by a qualified phonetician. It is the intention that if any researchers produce what are verified as more accurate transcriptions, that they will also share these with the research community. The dataset may also prove of interest to researchers in speaker verification and other related applications. The authors hope that the baselines reported here will very quickly be exceeded by others.

XI. Conclusion

This paper has presented the TCD-TIMIT database, a labelled database of continuous speech for audio-visual speech recognition. It consists of 13826 video clips in MP4 format, yielding high-quality audio and video footage of 62 speakers reading a total of 6913 sentences from the TIMIT corpus. Each sentence was recorded from two camera angles; straight and 30 degrees to the speaker's right. Three of the speakers are professional lipspeakers. Results of the numerous baseline experiments run on the database have been reported here to allow researchers progress the state of the art in audio-visual speech recognition research. In particular, it is hoped that comparative studies of lipspeaker and volunteer performance will further improve the visual feature extraction used in the front ends of AVSR systems in the future.

References

[1] E. D. Petajan, "Automatic Lipreading to Enhance Speech Recognition (Speech Reading)," Ph.D. dissertation, Univ. Illinois Urbana-Champaign, Champaign, IL, USA, 1984.

[2] G. Potamianos, C. Neti, G. Gravier, A. Garg, and A. Senior, "Recent advances in the automatic recognition of audiovisual speech," in *Proc. IEEE*, Sep. 2003, vol. 91, no. 9, pp. 1306–1326.

[3] C. Chibelushi, F. Deravi, and J. S. D. Mason, "A review of speech-based bimodal recognition," *IEEE Trans. Multimedia*, vol. 4, no. 1, pp. 23–37, Mar. 2002.

[4] R. Goecke, J. B. Millar, A. Zelinsky, and J. Robert-Ribes, "A detailed description of the Avozes data corpus," in *Proc. 10th Aust. Int. Conf. Speech Sci. Technol.*, 2004, pp. 486–491.

[5] C. Sui, S. Haque, R. Togneri, and M. Bennamoun, "A 3D audio-visual corpus for speech recognition," in *Proc. SST*, 2012, pp. 125–128.

[6] T. Gan, "Bimodal speech recognition," Ph.D. dissertation, Dept. Informat., Hamburg Univ., Hamburg, Germany, 2012 [Online]. Available: http://ediss.sub.uni-hamburg.de/volltexte/2012/5946

[7] L. Cappelletta, "What is a viseme? Exploring the visual side of automatic speech recognition," M.S. thesis, Dept. Electron. Electr. Eng., Trinity College Dublin, Dublin, Ireland, 2012.

[8] A. Pass, J. Zhang, and D. Stewart, "An investigation into features for multi-view lipreading," in *Proc. 17th IEEE Int. Conf. Image Process.*, Sep. 2010, pp. 2417–2420.

[9] A. G. Chitu and L. Rothkrantz, "Building a data corpus for audio-visual speech recognition," in *Proc. Euromedia 2007*, Apr. 2007, pp. 88–92.

[10] L. F. Lamel, R. Kassel, and S. Seneff, "Speech database development: Design and analysis of the acoustic-phonetic corpus," in *Proc. DARPA Speech Recog. Workshop*, 1986, pp. 100–110.

[11] K. Driel, "Building a visual speech recognizer," M.S. thesis, Faculty Electr. Eng., Math., Comput. Sci., Delft Univ. Technol., Delft, The Netherlands, 2009.

[12] X. Lin, H. Yao, X. Hong, and Q. Wang, "HIT-AVDB-II: A new multi-view and extreme feature cases contained audio-visual database for biometrics," in *Proc. 11th Joint Conf. Inf. Sci.*, , 2008, pp. 1–7.

[13] T. Chen, "Audiovisual speech processing," *IEEE Signal Process. Mag.*, vol. 18, no. 1, pp. 9–21, Jan. 2001.

[14] I. Matthews, T. Cootes, J. Bangham, S. Cox, and R. Harvey, "Extraction of visual features for lipreading," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 24, no. 2, pp. 198–213, Feb. 2002.

[15] T. J. Hazen, K. Saenko, C.-H. La, and J. R. Glass, "A segment-based audio-visual speech recognizer: Data collection, development, and initial experiments," in *Proc. 6th Int. Conf. Multimodal Interfaces*, New York, NY, USA, 2004, pp. 235–242.

[16] T. Kleinschmidt, D. Dean, S. Sridharan, and M. Mason, "A continuous speech recognition evaluation protocol for the AVICAR database," in *Proc. Int. Conf. Signal Process. Commun. Syst.*, Gold Coast, Australia, 2007, pp. 339–345.

[17] E. Bailly-baillire *et al.*, "The BANCA database and evaluation protocol," in *Proc. Int. Conf. Audio- Video-Based Biometric Person Authentication*, 2003, pp. 625–638.

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

[19] C. C. Chibelushi, S. Gandon, J. S. D. Mason, F. Deravi, and R. D. Johnston, "Design issues for a digital audio-visual integrated database," in *Proc. IEE Colloq. Integr. Audio-Visual Process. Recog., Synthesis Commun.*, Nov. 1996, pp. 7/1–7/7.

[20] M. Cooke, J. Barker, S. Cunningham, and X. Shao, "An audio-visual corpus for speech perception and automatic speech recognition," *J. Acoust. Soc. Amer.*, vol. 120, p. 2421, 2006.

[21] I. Matthews, G. Potamianos, C. Neti, and J. Luettin, "A comparison of model and transform-based visual features for audio-visual LVCSR," in *Proc. Int. Conf. Multimedia Expo*, 2001, pp. 22–25.

[22] C. Sanderson, *Biometric Person Recognition: Face, Speech and Fusion*. Saarbrücken, Germany: VDM, 2008.

[23] N. A. Fox, B. A. O'Mullane, and R. B. Reilly, "VALID: Anew practical audio-visual database, and comparative results.," in *Audio-and Video-Based Biometric Person Authentication*. Berlin, Germany: Springer-Verlag, 2005.

[24] J. Movellan, , G. Tesauro, D. Touretzky, T. Leen, and S. Mateo, Eds., "Visual Speech Recognition with Stochastic Networks," in *Advances in Neural Information Processing Systems*. Cambridge, MA, USA: MIT Press, 1995, vol. 7, pp. 851–858.

[25] K. Messer, J. Matas, J. Kittler, and K. Jonsson, "XM2VTSDB: The extended M2VTS database," in *Proc. 2nd Int. Conf. Audio- Video-Based Biometric Person Authentication*, 1999, pp. 72–77.

[26] K. Kumar, T. Chen, and R. Stern, "Profile view lip reading," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, Apr. 2007, vol. 4, pp. IV-429–IV-432.

[27] Y. Lan, B.-J. Theobald, and R. Harvey, "View independent computer lip-reading," in *Proc. 2012 IEEE Int. Conf. Multimedia Expo*, Washington, DC, USA, 2012, pp. 432–437.

[28] G. Galatas, G. Potamianos, and F. Makedon, "Audiovisual speech recognition incorporating facial depth information captured by the Kinect," in *Proc. 20th Eur. Signal Process. Conf.*, 2012, pp. 2714–2717.

[29] Y. W. Wong *et al.*, "A new multi-purpose audio-visual UNMC-VIER database with multiple variabilities," *Pattern Recog. Lett.*, vol. 32, no. 13, pp. 1503–1510, 2011.

[30] D. Burnham *et al.*, "A blueprint for a comprehensive Australian English auditory-visual speech corpus," in *Sel. Proc. 2008 HCSNet Workshop Designing Aust. Nat. Corpus*, 2009, pp. 96–107.

[31] J. S. Garofolo *et al.*, DARPA TIMIT Acoustic Phonetic Continuous Speech Corpus CDROM. 1993 [Online]. Available: https://catalog.ldc.upenn.edu/LDC93S1, Accessed on: Mar. 3, 2015

[32] T. Saitoh and R. Konishi, "A study of influence of word lip reading by change of frame rate," in *Proc. AVSP-2010*, 2010, pp. 131–136.

[33] P. Lucey and G. Potamianos, "Lipreading using profile versus frontal views," in *Proc. IEEE 8th Workshop Multimedia Signal Process.*, Oct. 2006, pp. 24–28.

[34] *Algorithms to Measure Audio Programme Loudness and True-Peak Audio Level*, I. R. BS1770-3, 2010 [Online]. Available: http://www.itu.int/rec/R-REC-BS.1770

[35] F. Brugnara, D. Falavigna, and M. Omologo, "Automatic segmentation and labeling of speech based on hidden Markov models," *Speech Commun.*, vol. 12, no. 4, pp. 357–370.

[36] B. L. Pellom and J. H. Hansen, "Automatic segmentation and labeling of speech recorded in unknown noisy channel environments," *Speech Comm.*, vol. 25, no. 1, pp. 97–116, Nov. 1998.

[37] K. Sjölander, "An hmm-based system for automatic segmentation and alignment of speech," in *Proc. Fonetik*, 2003, pp. 93–96.

[38] K. Vertanen, "Baseline WSJ acoustic models for HTK and Sphinx: Training recipes and recognition experiments," Cavendish Lab., Univ. Cambridge, Cambridge, U.K., Rep., 2006 [Online]. Available: http://westray.eng.cam.ac.uk/is/papers/baseline_wsj_recipes.pdf, Accessed on: Mar. 3, 2015

[39] J.-P. Hosom, "Automatic time alignment of phonemes using acoustic-phonetic information," Ph.D. dissertation, Oregon Graduate Inst. Sci. Technol., Beaverton, OR, USA, 2000.

[40] J. Yuan and M. Liberman, "Speaker identification on the SCOTUS corpus," *J. Acoust. Soc. Amer.*, vol. 123, no. 5, p. 3878, 2008.

[41] K.-F. Lee and H.-W. Hon, "Speaker-independent phone recognition using hidden Markov models," *IEEE Trans. Acoust., Speech Signal Process.*, vol. 37, no. 11, pp. 1641–1648, Nov. 1989.

[42] A. J. Bliss, "Languages in contact: Some problems of hiberno-english," in *Proc. Royal Irish Academy. Section C: Archaeol., Celtic Studies, History, Linguistics, Literature*, 1972, pp. 63–82.

[43] J. Jeffers and M. Barley, *Speechreading: (Lipreading)*. Springfield, IL, USA: Charles C. Thomas, 1971.

[44] S. J. Young *et al., The HTK Book, Version 3.4*. Cambridge, U.K.: Cambridge Univ. Press, 2006.

[45] M. Heckmann, K. Kroschel, C. Savariaux, and F. Berthommier, "DCT-based video features for audio-visual speech recognition," in *Proc. INTERSPEECH*, 2002, pp. 1925–1928.

[46] R. Seymour, D. Stewart, and J. Ming, "Comparison of image transform-based features for visual speech recognition in clean and corrupted videos," *J. Image Video Process.*, vol. 2008, pp. 14:1–14:9, Jan. 2008.

[47] P. Scanlon, R. Reilly, and P. d. Chazal, "Visual feature analysis for automatic speechreading," in *Proc. Int. Conf. Audio-Visual Speech Process.*, 2003, pp. 127–132.

[48] C. Neti *et al.*, "Audio-visual speech recognition," CLSP Summer Workshop, Johns-Hopkins Univ., Baltimore, MD, USA, Rep. EPFL-REPORT-82633, 2000.

[49] X. Zhang, C. C. Broun, R. M. Mersereau, and M. A. Clements, "Automatic speechreading with applications to human-computer interfaces," *EURASIP J. Appl. Signal Process.*, vol. 2002, no. 1, pp. 1228–1247, Jan. 2002.

[50] J. Gowdy, A. Subramanya, C. Bartels, and J. Bilmes, "DBN based multi-stream models for audio-visual speech recognition," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, May 2004, vol. 1, p. I-993-6.

[51] G. Galatas, G. Potamianos, A. Papangelis, and F. Makedon, "Audio visual speech recognition in noisy visual environments," in *Proc. 4th Int. Conf. Pervasive Technol. Related Assistive Environ.*, New York, NY, USA, 2011, pp. 19:1–19:4.

[52] G. Papandreou, A. Katsamanis, V. Pitsikalis, and P. Maragos, "Adaptive multimodal fusion by uncertainty compensation with application to audiovisual speech recognition," *IEEE Trans. Audio, Speech, Language Process.*, vol. 17, no. 3, pp. 423–435, Mar. 2009.

[53] G. Potamianos, J. Luettin, and C. Neti, "Hierarchical discriminant features for audio-visual LVCSR," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, May 2001, vol. 1, pp. 165–168.

[54] S. Taylor, B.-J. Theobald, and I. Matthews, "The effect of speaking rate on audio and visual speech," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, May 2014, pp. 3037–3041.

Naomi Harte (M'05) is an Assistant Professor with the School of Engineering, Trinity College Dublin, Dublin, Ireland, where she was appointed as an SFI Engineering Initiative Lecturer in digital media in 2008. Prior to returning to academia, she worked in high-tech start-ups in the field of DSP systems development. Her research interests include speech quality, audio-visual speech recognition, emotion in speech, speaker verification, and bird species analysis from song.

Eoin Gillen received the B.Sc. degree in computer and electronic engineering and the M.Sc. degree in audio-visual speech recognition from Trinity College Dublin, Dublin, Ireland, in 2012 and 2013, respectively.

Since 2013, he has been a Research Assistant with the Sigmedia Group, School of Engineering, Trinity College Dublin, Dublin, Ireland. His research interests include audio-visual media production and manipulation, software development, developing multimedia datasets, and conducting large-scale MOS and MUSHRA tests.