LETTERS TO THE EDITOR

This Letters section is for publishing (a) brief acoustical research or applied acoustical reports, (b) comments on articles or letters previously published in this Journal, and (c) a reply by the article author to criticism by the Letter author in (b). Extensive reports should be submitted as articles, not in a letter series. Letters are peer-reviewed on the same basis as articles, but usually require less review time before acceptance. Letters cannot exceed four printed pages (approximately 3000–4000 words) including figures, tables, references, and a required abstract of about 100 words.

An audio-visual corpus for speech perception and automatic speech recognition (L)

Martin Cooke[a] and Jon Barker
Department of Computer Science, University of Sheffield, Regent Court, 211 Portobello Street, Sheffield, S1 4DP United Kingdom

Stuart Cunningham
Department of Human Communication Sciences, University of Sheffield, Sheffield, S1 4DP United Kingdom

Xu Shao
Department of Computer Science, University of Sheffield, Regent Court, 211 Portobello Street, Sheffield, S1 4DP United Kingdom

(Received 29 November 2005; revised 2 April 2006; accepted 26 June 2006)

An audio-visual corpus has been collected to support the use of common material in speech perception and automatic speech recognition studies. The corpus consists of high-quality audio and video recordings of 1000 sentences spoken by each of 34 talkers. Sentences are simple, syntactically identical phrases such as "place green at B 4 now." Intelligibility tests using the audio signals suggest that the material is easily identifiable in quiet and low levels of stationary noise. The annotated corpus is available on the web for research use.
[DOI: 10.1121/1.2229005]

PACS number(s): 43.71.Es, 43.72.Ne, 43.66.Yw [DOS] Pages: 2421–2424

I. INTRODUCTION

Understanding how humans process and interpret speech in adverse conditions is a major scientific challenge. Two distinct methods for modeling speech perception have been studied. The traditional approach has been to construct "macroscopic" models, which predict overall speech intelligibility in conditions of masking and reverberation. Models such as the articulation index (French and Steinberg, 1947), the speech transmission index (Steeneken and Houtgast, 1980), and the speech intelligibility index (ANSI S3.5, 1997) fall into this category. A more recent idea is to apply automatic speech recognition (ASR) technology to construct what might be called "microscopic" models of speech perception, which differ from macroscopic approaches in their additional capability to predict listeners' responses to individual tokens. Examples of microscopic models include Ghitza (1993), Ainsworth and Meyer (1994), Holube and Kollmeier (1996), and Cooke (2006).

Although microscopic modeling results have been promising, a serious barrier to further development of these models has been the lack of suitable speech material. Unlike speech perception studies, microscopic models require a large volume of speech material for training purposes. Many corpora for ASR exist, but the use of such corpora in speech perception testing is problematic. Speech material tends to be uncontrolled, phonetically unbalanced, or consists of tokens whose durations make them unsuitable for behavioral studies. By contrast, corpora used in perceptual studies tend to be too small or insufficiently varied for microscopic models of speech perception.

Previous models have attempted to explain the *auditory* perception of speech signals. However, speech production results in both acoustic and optical signals. It has become increasingly clear that the visual modality has a fundamental role in speech perception, and any full perceptual account needs to explain the complicated interactions between modalities (Rosenblum, 2002). Acoustically confusable phoneme pairs such as /m/ and /n/ can be disambiguated using visual cues. Automatic speech recognition systems can exploit these cues to improve audio-only recognition performance in both clean and noisy conditions (Potamianos *et al.*, 2003). Visual cues can also be used to separate speech from competing noise sources. One particularly interesting area of study in this respect is the audio-visual separation of simultaneous cochannel speech. Despite the clear importance of visual speech information, until now there have been no eas-

[a]Electronic mail: m.cooke@dcs.shef.ac.uk

ily accessible corpora suitable for building multimodal models. However, recent advances in video compression technology, the rapidly falling cost of hard disk storage, the increasing capacity of optical storage media, and the increasing bandwidth of typical Internet connections, mean that storage and distribution are no longer a barrier.

These factors motivated the collection of an audio-visual corpus designed for both ASR-based and perceptual studies of speech processing. The form of the corpus was heavily influenced by the coordinate response measure (CRM) task (Moore, 1981; Bolia *et al.*, 2000), which consists of simple sentences of the form "READY <call sign> GO TO <color> <digit> NOW." CRM used 8 call signs, 4 colors, and 8 digits and each combination was spoken once by 8 talkers for a total of 2048 sentences. CRM is useful for studies of early processes in speech perception since it contains sentence length material, yet is devoid of high-level linguistic cues. The design of CRM makes it valuable in multitalker tasks (e.g., Brungart *et al.*, 2001) where listeners are asked to identify the color-digit combination spoken by the talker who provided a given call sign.

The new collection, which we call the Grid corpus, consists of sentences such as "place blue at F 9 now" of the form <command:4> <color:4> <preposition:4> <letter:25> <digit:10> <adverb:4>," where the number of choices for each component is indicated. Grid extends CRM in a number of ways. The set of talkers is larger (34 rather than 8) and the number of sentences per talker is 1000 rather than 256, giving a total corpus size of 34 000 as opposed to 2048 sentences. Consequently, Grid contains greater variety and is large enough to meet the training requirements of ASR systems. Grid has an improved phonetic balance due to the use of alphabetic letters, which also presents listeners with a more difficult task than the four color options of CRM. Grid is more varied than CRM since the "filler" items (command, preposition, and adverb) are no longer static. This also prevents echo-like artifacts arising when two or more sentences with identical fillers are summed in, for example, experiments involving multiple simultaneous talkers. Finally, Grid provides speech video as well as audio, allowing the development of multimodal perceptual models.

While the primary motivation for the Grid corpus was to support the construction of microscopic, multimodal models of speech perception, it can also be used for conventional behavioral studies of audio and audio-visual speech perception. Similarly, Grid is valuable for ASR studies of speech in noise, the separation of speech from multitalker backgrounds, and audio-visual speech recognition and separation.

II. CORPUS

A. Sentence design

Each sentence consisted of a six word sequence of the form indicated in Table I. Of the six components, three—color, letter, and digit—were designated as "keywords". In the letter position, "w" was excluded since it is the only multisyllabic English alphabetic letter. "Zero" was used rather than "oh" or "nought" to avoid multiple pronunciation alternatives for orthographic "0." Each talker produced all combinations of the three keywords, leading to a total of 1000 sentences per talker. The remaining components—command, preposition, and adverb—were "fillers." Four alternatives were available in each filler position. Filler words were chosen to create some variation in contexts for the neighboring key words. Different gross phonetic classes (nasal, vowel, fricative, plosive, liquid) were used as the initial or final sounds of filler words in each position.

TABLE I. Sentence structure for the Grid corpus. Keywords are identified with asterisks.

command	color*	preposition	letter*	digit*	adverb
bin	blue	at	A–Z	1–9, zero	again
lay	green	by	excluding W		now
place	red	in			please
set	white	with			soon

B. Speaker population

The aim of speaker selection was to provide a sufficiently large number of speakers to allow users of the corpus to select subsets based on criteria such as intelligibility, homogeneity, and variety. Sixteen female and 18 male talkers contributed to the corpus. Participants were staff and students in the Departments of Computer Science and Human Communication Science at the University of Sheffield. Student participants were paid for their contribution. All spoke English as their first language. All but three participants had spent most of their lives in England and together encompassed a range of English accents. Two participants grew up in Scotland and one was born in Jamaica. Ages ranged from 18 to 49 years (mean: 27.4 years).

C. Collection

Audio-visual recordings were made in an IAC single-walled acoustically isolated booth. Speech material was collected from a single Bruel & Kjaer (B & K) type 4190 $\frac{1}{2}$-in. microphone placed 30 cm in front of the talker. The signal was preamplified by a B & K Nexus model 2690 conditioning amplifier prior to digitization at 50 kHz by a Tucker-Davis Technologies System 3 RP2.1 processor. Collection of speech material was under computer control. Sentences were presented on a computer screen located outside the booth, and talkers had 3 s to produce each sentence. Talkers were instructed to speak in a natural style. To avoid overly careful and drawn-out utterances, they were asked to speak sufficiently quickly to fit into the 3-s time window. Talkers were allowed to repeat the sentence if they felt it necessary, either because of a mistake during production or if part of the utterance fell outside the 3-s window. As an aid, the captured waveform was displayed on the screen. In addition, talkers were asked to repeat the utterance if the captured waveform was judged by the software to be too quiet or too loud. Prior to saving, signals were scaled so that the maximum absolute value was unity, in order to optimize the use of the quantized amplitude range. Scale factors were stored to allow the normalization process to be reversed.

A simultaneous continuous video recording was made on to MiniDV tape using a Canon XM2 video camcorder. The camera was set up to capture full frames at 25 frames/s.

To avoid both noise and distraction from the video apparatus, the camera was placed at eye level outside the booth, abutting the booth window. Light sources were arranged to produce uniform illumination across the face, and the subject was seated in front of a plain blue background. To ease temporal alignment of the audio and visual signals, the camera took its audio input from the high-quality audio signal provided by the microphone in the booth.

Although talkers were allowed to repeat an utterance if they misread the prompt, they occasionally made errors without realizing they had done so. A semi-automatic screening procedure was employed to locate errors in the corpus. The screening process used an ASR system based on talker-dependent, whole-word hidden Markov models (HMM), trained separately for each talker. A "jack-knife" training procedure was used in which 80% of the talker's utterances were used for training and the remaining 20% recognized. This procedure was performed five times with a different subset of utterances so that the subset recognized was independent of the training set on each occasion. The ASR system produced word-level transcripts for each utterance, and errors were flagged if the recognition output differed from the sentence the talker was meant to have read. On average, 57 out of 1000 utterances were flagged per talker. Flagged utterances were checked by a listener and any sentences with errors were marked for re-recording. Talkers were recalled to perform the re-recording session, during which time their utterances were monitored over headphones. Talkers were asked to repeat any incorrectly produced utterances. In all, 640 utterances (an average of 19 per talker or 1.9% of the corpus) were re-recorded.

While the screening process guaranteed that many of the errors in the corpus were corrected, it is possible that some errors were not detected. For a spoken sentence containing an error to appear correct, the recognition system must have made a complementary error (i.e., an error which corrects the error made by the talker). However, since the error rates of both the talkers and recognizer are very low, the conjunction of complementary errors is extremely unlikely. Informal human screening of a subset of utterances led to an estimate of an error rate of not more than 0.1% (i.e., one error per 1000 utterances). Most of the errors detected involved misproduced filler items, so the number of sentences containing misproduced keywords is smaller still.

D. Postprocessing

1. Audio

Prior to further processing, audio signals were downsampled to 25 kHz using the MATLAB *resample* routine. A subset of 136 utterances (four randomly chosen from each talker) was used to estimate the peak S/N according to the ITU P.56 standard (ITU-T, 1993). The peak S/N varied across talkers from 44 to 58 dB (mean=51 dB, s.d.=3.6 dB).

The talker-dependent HMM-based ASR systems used in the screening of speaker errors were employed to estimate the alignment between the word-level transcription and the utterance. In addition, phone-level transcriptions of each utterance were produced by forced alignment using the HVITE program in the HTK hidden Markov model toolkit (Young *et al.*, 1999). Pronunciations were taken from the British English Example Pronunciation dictionary (BEEP[1]). To bootstrap the initial set of HMMs, 60 sentences from one of the speakers were manually transcribed at the phone level.

2. Video

Unlike the audio collection, which was computer-controlled, the video data were collected continuously throughout the recording session, and thus contained both genuine Grid utterances as well as false starts, incorrect utterances, and other material. Consequently, it was necessary to extract video segments corresponding to the final endpointed audio recordings. Utterance segments were first located approximately using a timestamp recorded by the software controlling the audio recording session. A precise location was then found by searching in this region for the 3-s period of the video file that best correlated with the 3 s of high-quality audio captured by the TDT processor. Correlations were performed using the smoothed energy envelope of the signals. Once the audio was precisely located, the corresponding 75-frame (i.e., 3-s) segment of video was extracted. The high-quality 50-kHz audio captured by the TDT processor was resampled to 44.1 kHz and used to replace the audio track of the video segment.

The DV format video was converted to MPEG-1 format using FFMPEG.[2] Two compression rates were used to produce both high and moderate quality versions of the video data. The high-quality video employed a bandwidth of 6 Mbits per s (comparable to DVD quality), while the moderate quality version used a bandwidth of 600 Kbits per s (a quality intermediate between a typical business-oriented videoconferencing system and VHS video). In both cases, the audio bit rate was set to 256 kbits per s.

III. AUDIO INTELLIGIBILITY TESTS

Twenty listeners with normal hearing heard independent sets of 100 sentences drawn at random from the corpus. All speech material had initial and trailing silence removed prior to presentation using utterance endpoints derived from the word alignments. Utterances were scaled to produce a presentation level of approximately 68 dB SPL and were presented diotically over Sennheiser HD250 headphones in the IAC booth. Listeners were asked to identify the color, letter, and digit spoken and entered their results using a conventional computer keyboard in which four of the nonletter/digit keys were marked with colored stickers. Those keys representing colors were activated immediately following the onset of each utterance. As soon as a color key was pressed, the 25 relevant letter keys were enabled, followed by the 10 digit keys. This approach allowed for rapid and accurate data entry: most listeners were able to identify a block of 100 utterances in 5–7 min. Listeners were familiarized with the stimuli and the task by identifying an independent practice set of 100 sentences prior to the main set.

Figure 1 (triangles) shows the mean scores and their standard errors across listeners. Unsurprisingly, fewer errors were made for colors (0.25% of the 2000 sentences) than for digits (0.7%) or letters (0.95%). At least one error occurred

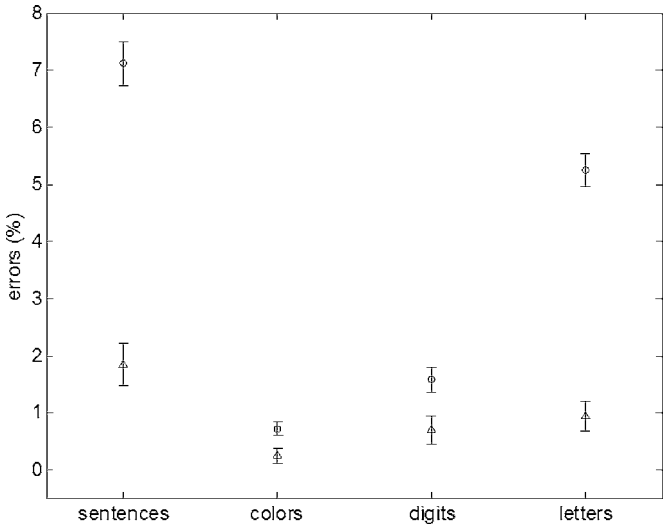

FIG. 1. Mean error rates across listeners for sentences, colors, digits, and letters. A sentence contained an error if one or more keywords were incorrectly identified. Triangles: clean sentence material; circles: sentences in speech-shaped noise. Error bars denote +/−1 standard errors.

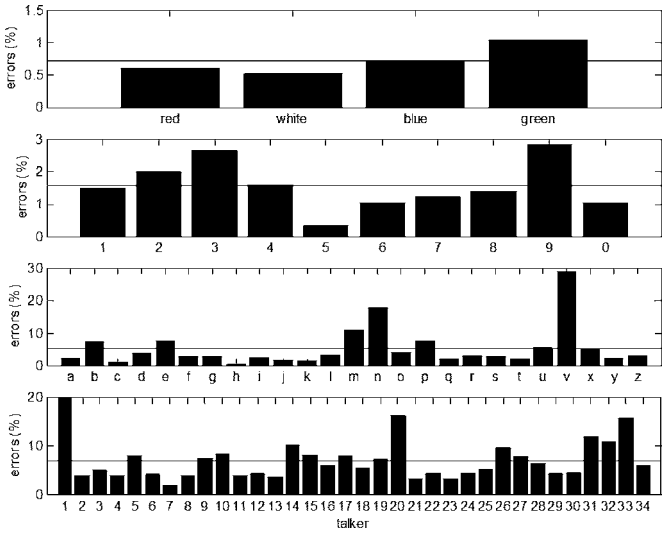

FIG. 2. Error rates for colors, digits, letters, and per talker for the noise conditions. Horizontal lines indicate mean error rates in each category. Talker error rates are measured as percentages of utterances in which at least one keyword was misidentified.

in 1.85% (37 out of 2000) of sentences. These low error rates suggest that the speech material collected was of high intelligibility overall.

Insufficient errors were reported using clean speech material to allow a more detailed inspection of the intelligibility of individual keywords or talkers. To support such an analysis, the same 20 listeners heard three further independent sets of 100 utterances mixed with speech-shaped noise whose spectrum matched the long-term spectrum of the Grid corpus at three signal-to-noise ratios: 6, 4, and 2 dB, producing a total of 6000 responses. Figure 1 (circles) shows error rates for colors (0.7%), digits (1.6%), letters (5.2%), and wholly correct sentences (7.1%). Figure 2 depicts the distribution of errors by keyword and across talkers. While the color and number distributions are reasonably flat, certain letters are recognized significantly less well than others. Inspection of letter confusion matrices revealed that most of the /v/ errors were caused by misidentification as /b/, while /m/ and /n/ tokens were confused with each other.

A range of identification rates (defined as the percentage of utterances in which at least one keyword was misidentified) across the 34 contributing talkers was observed. In particular, listeners misidentified keywords in utterances by talkers 1 (19.8%), 20 (16.2%), and 33 (15.6%), while fewer than 2% of sentences spoken by talker 7 were misidentified by this listener group. However, most talkers produced errors rates of around 5%.

IV. SUMMARY

Grid, a large multitalker audio-visual sentence corpus, has been collected to support joint computational-behavioral studies in speech perception. Audio-only intelligibility tests suggest that the speech material is easily identified in quiet and low-noise conditions. Further tests of visual and audio-visual intelligibility are planned. The complete corpus and transcriptions are freely available for research use at the website <http://www.dcs.shef.ac.uk/spandh/gridcorpus>.

ACKNOWLEDGMENTS

Corpus collection and annotation was supported by grants from the University of Sheffield Research Fund and the UK Engineering and Physical Research Council (GR/T04823/01). The authors thank Dr. Anthony Watkins and two anonymous reviewers for helpful comments.

[1]Available at ftp://svr-ftp.eng.cam.ac.uk/pub/comp.speech/dictionaries/
[2]Available at http://ffmpeg.sourceforge.net/index.php

Ainsworth, W. A., and Meyer, G. F. (**1994**). "Recognition of plosive syllables in noise: Comparison of an auditory model with human performance," J. Acoust. Soc. Am. **96**, 687–694.

ANSI (**1997**). ANSI S3.5-1997, "American National Standard Methods for the Calculation of the Speech Intelligibility Index" (American National Standards Institute, New York).

Bolia, R. S., Nelson, W. T., Ericson, M. A., and Simpson, B. D. (**2000**). "A speech corpus for multitalker communications research," J. Acoust. Soc. Am. **107**, 1065–1066.

Brungart, D. S., Simpson, B. D., Ericson, M. A., and Scott, K. R. (**2001**). "Informational and energetic masking effects in the perception of multiple simultaneous talkers," J. Acoust. Soc. Am. **100**, 2527–2538.

Cooke, M. P. (**2006**). "A glimpsing model of speech perception in noise," J. Acoust. Soc. Am. **119**, 1562–1573.

French, N., and Steinberg, J. (**1947**). "Factors governing the intelligibility of speech sounds," J. Acoust. Soc. Am. **19**, 90–119.

Ghitza, O. (**1993**). "Adequacy of auditory models to predict human internal representation of speech sounds," J. Acoust. Soc. Am. **93**, 2160–2171.

Holube, I., and Kollmeier, B. (**1996**). "Speech intelligibility prediction in hearing-impaired listeners based on a psychoacoustically motivated perception model," J. Acoust. Soc. Am. **100**, 1703–1716.

ITU-T (**1993**). "Objective measurement of active speech level," ITU-T Recommendation P. 56.

Moore, T. (**1981**). "Voice communication jamming research," in *AGARD Conference Proceedings 331: Aural Communication in Aviation*, Neuilly-Sur-Seine, France, 2:1–2:6.

Potamianos, G., Neti, C., Gravier, G., Garg, A., and Senior, A. W. (**2003**). "Recent advances in the automatic recognition of audiovisual speech," Proc. IEEE **91**, 1306–1326.

Rosenblum, L. D. (**2002**). "The perceptual basis for audiovisual integration," in *Proceedings International Conference on Spoken Language Processing*, 1461–1464.

Steeneken, H., and Houtgast, T. (**1980**). "A physical method for measuring speech-transmission quality," J. Acoust. Soc. Am. **67**, 318–326.

Young, S., Kershaw, D., Odell, J., Ollason, D., Valtchev, V., and Woodland, P. (**1999**). *The HTK Book 2.2* (Entropic, Cambridge).