Speech Perception and Automated Speech Recognition Using an Audio-Visual Corpus

SHUBHAM GUPTA 1, DR. ALOK SEMWAL 2, DR. ABHILASH SINGH 3

1Department of Computer Science & Engineering, Shivalik College of Engineering, Dehradun
2College of Pharmacy, Shivalik, Dehradun
3Shivalik Institute of Professional Studies, Dehradun

shubham.gupta@sce.org.in

ABSTRACT: To make it simpler to use among that in speech perception and automatic voice recognition investigations, an audio-visual corpus has been constructed. Each of the 34 talkers spoke 1000 words in high-quality audio and video recordings that make up the corpus. Short, syntax connected phrases like "put green at B 4 immediately" are included in sentences. Audio signals used in intelligence tests reveal that the substance is clearly recognisable in still, low levels of stationary noise. The annotation corpus is accessible for study over the internet. A crucial area of science is understanding how people process information and make sense of it in difficult conditions. The researchers looked at two approaches to replicate speech perception. Traditionally, total speech intelligibility has been predicted using "macroscopic" models that account for reverberation and masking. The Steeneken and Houtgast voice propagation index from 1980, the French and Steinberg articulation index from 1947, and the ANSI S3.5 speech intelligibility index from 1997 are among the models in this collection. A relatively recent notion is to create as such "microscopic" models of speech perception using automated speech recognition (ASR) technology. These models vary from macroscopic models in that they may anticipate listeners' reactions to particular tokens.

KEYWORDS: Automatic, Audio-Visual, Corpus, Internet, Speech Recognition.

1. INTRODUCTION

Although the findings of microscopic modeling have been encouraging, a major stumbling block to continued development of these models has been a lack of appropriate speech material. Microscopic models need a substantial quantity of speech data for training, in contrast to speech perception research. Although there are many ASR corpora available, it is challenging to use them in speech perception tests. For behavioural study, speech material is generally uncontrolled, phonetically unbalanced, or composed of tokens with brief durations. For microscopic models of speech perception, however, corpora employed in perceptual research are often too tiny or insufficiently diverse. Previous models have tried to describe how speech signals are perceived by the ears. Speech, on the other hand, generates both acoustic and visual impulses [1].

Speech perception increasingly depends on the visual modality, and any comprehensive perceptual explanation must account for the intricate interactions between the various modalities. Rosenblum, 2002. Visual cues may be used to distinguish morpheme pairs like /m/ and /n/ that are auditory ambiguous. These signals may be used by automatic speech recognition systems to enhance sound recognition performance in both calm and noisy environments. Potamianos and many others, 2003 Speech may be distinguished from rival sound sources using visual signals as well. In this perspective, a particularly exciting field of research is the audio-visual separation of control speech [1]. Despite the apparent value of visual speech data, there has never been a simple method to record it. The terms of the Acoustical Society of America's (ASA) licence or copyright are applicable to distribution; see corpora that are readily accessible and may be utilised to build multimodal algorithms. As according current developments in video compression technologies, the fast declining cost of hdd storage, the increasing capacity of optical storage devices, and the increasing speed of conventional Internet connections, storing and transmitting are no longer problems. These considerations led to the development of an audio-visual corpus that could be utilized for both ASR and perceptual investigations of speech perception. metric for a coordinated response [2] [3]. Grid has a greater phonological balance than CRM's four colour possibilities, although hearing is more difficult due to the usage of alphabetic letters. Since the "filler" elements command, conjunction, and adverb are no longer static, Grid is more diverse than CRM. This also precludes the development of echo-like effects when three or even more sentences with equivalent fillers are merged, as in tests with several simultaneous talkers. Finally, Grid incorporates both audio and video to assist the construction of multimodal perception models.

The Grid dataset may be used to create microscopic, multimodal models of word recognition in addition to typical behavioural studies on audio and sound speech perception. Grid is useful for ASR analyses of speech in noisy contexts, speech distinction from multitalker backgrounds, and sound speech recognition and isolation. A. Each phrase has a six-word triad as its building block. Color, letter, and number were three of the six components that were categorised as "keywords". Considering that the letter "w" is the only multisyllabic English alphabetic letter, it was omitted from the letter position. To remove many possible pronunciations for the orthographic "0," "zero" [4] was chosen instead of "oh" or "nought." Each talker produced 1000 sentences overall by using all three keyword combinations. The command, preposition, and adverb were "fillers," as were the other components. There were four alternatives for each filling place.

1.1. Number of native speakers

A significant number of speakers required to be made available in order for corpus users to choose subsets based on factors like intelligibility, homogeneity, etc variety. The corpus consisted of 16 female talkers and 18 male talkers. Staff members and students from the University of Sheffield's departments of computer humans and scientific communications science participated in the study. Students earned compensated for participating. They could all speak English quite well. All but three of the participants had lived the most of their lives in England, and the participants' combined accents spanned a broad spectrum of English dialects. Two of the candidates were Jamaican citizens by birth but Scottish citizens by birth. The age ranged from 18 to 49 years, with a mean of 27.4 years. The essence of the speech was recorded using computer technology. On a computer screen outside the booth, talkers were given three seconds to complete their phrase. Talkers were told to use informal language. In order to avoid being excessively cautious and making long comments, they were instructed to talk as rapidly as possible to keep under the 3-second time restriction. If a production problem or a section of the speech extended more than the permitted three seconds, talkers had the option to restart the statement.

The recorded waveform was shown on the screen as a visual assistance. Talkers were also prompted to repeat their utterances if the recorded waveform was deemed to be too quiet or too loud by the program. To make the most of the quantized amplitude range, signals were scaled prior to storing such that the highest absolute value was unity. To reverse the normalisation process, the scale factors were stored. A continual video recording on MiniDV tape was created using a Canon XM2 video camera. According to Figure 1, the camera was set up to capture full frames at a rate of 25 frames per second [5].

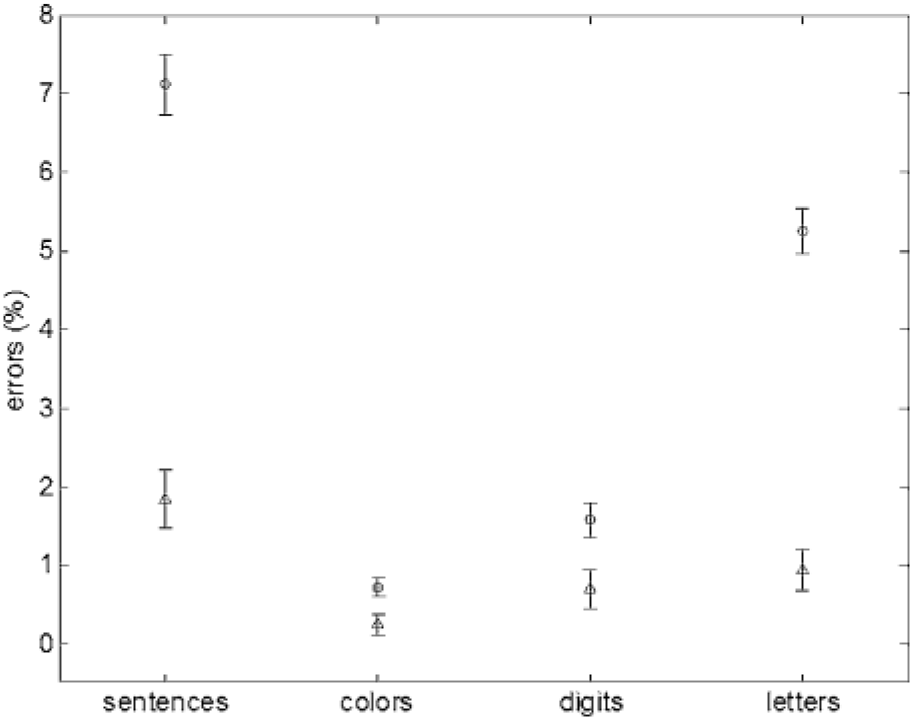

Figure 1: A Sentence Contained An Error. Triangles indicate clear sentences; circles indicate sentences with speech-like noise. Message Bars Indicate +/1 Standard Mistakes.

2. DISCUSSION

Despite the fact that talkers were permitted to repeat a speech if they misunderstood the cue, they sometimes committed mistakes without recognizing it. Semi-automated screening was employed to discover errors in the corpus. An ASR system built on whole-word HMMs that were individually trained for each talker was used during the screening phase. The "jack-knife" method of teaching was used, in which training was done on 80percent of the talker's utterances and identification on the remaining 20 percent. This process was repeated five times, using a new subset of words each time to make sure the subset that was discovered each time had nothing to with the training set. For each speech, the ASR system generated word-level transcripts, and when the output of recognition differed from the phrase the talker was intending to read, faults were flagged. An typical talker emphasised 57 out of every 1000 words they spoke. After the listener had evaluated the highlighted utterances, any sentences that included mistakes were marked for re-recording. The talkers were instructed to take part in a session of s new while their speech was being recorded and monitored via earphones. The speakers were asked to restate any false statements they had made. 640 utterances in all, or 1.9% of the sample, were revised [6].

While the screening procedure ensured that many of the mistakes in the corpus were rectified, some inaccuracies may have escaped detection. To make a spoken phrase with an error seem right, the recognition system must have produced a complimentary mistake, that is, an error that corrects the talker's error. However, since both the talkers' and the recognizer's mistake rates are very low, the occurrence of complimentary errors is highly improbable. Video Unlike the computer-controlled audio collection, video data were gathered constantly during the recording session, including both legitimate Grid utterances and false starts, erroneous utterances, and other stuff. As a result, video segments corresponding to the final end pointed audio recordings had to be extracted. The timestamp obtained by the program managing the audio recording session was used to find utterance fragments.

Twenty individuals with normal hearing were exposed to different sets of 100 phrases randomly selected from the corpus. Every piece of speech material had its beginning and finish quiet removed before being shown using utterance endpoints that were produced from word alignments [7]. Utterances were presented diametrically in the IAC booth while wearing Sennheiser HD250 headphones at a presentation level of around 68 dB SPL. A regular computer keyboard with colourful stickers on four of the no letter/digit buttons was used to let listeners to enter their replies after being asked to identify the colour, letter, and number spoken. Each beginning phrase activated the keys that depicted the colours. Following the activation of a colour key, the 10 number keys were followed by the 25 corresponding letter keys.

There weren't enough errors found in the clean speech material to allow for a full analysis of the understandability of certain talkers or keywords. To back up such a study, the same [8] 20 Three unique sets of 100 utterances each were presented to listeners, along with speech-shaped noise whose long-term spectrum matched that of the Grid corpus, at three different signal-to-noise ratios: 6, 4, and 2 dB, yielding a total of 6000 responses. Colors make mistakes at a rate of 0.7 percent, numerals at 1.6 percent, letters at 5.2 percent, and words that are 100 percent precise make mistakes at 7.1 percent. Figure 2 displays the distribution of errors by keywords and talkers. The distributions of colour and numbers are rather consistent, however certain letters are harder to recognise than others. According to letter confusion matrices, the bulk of /v/ errors were produced by mistaking them for /b/, while /m/ and /n/ tokens were mixed up with one another.

A variety of identification rates was found among the 34 contributing talkers, defined as the proportion of utterances in which at least one term was misdiagnosed. Listeners misunderstood terms in utterances by talkers 1 19.8 percent, 20 16.2 percent, and 33 15.6 percent, respectively, whereas this listener group misidentified less than 2% of sentences uttered by talker 7. The majority of talkers, on the other hand, had mistake rates of about 5% as shown in Figure 2 [9].

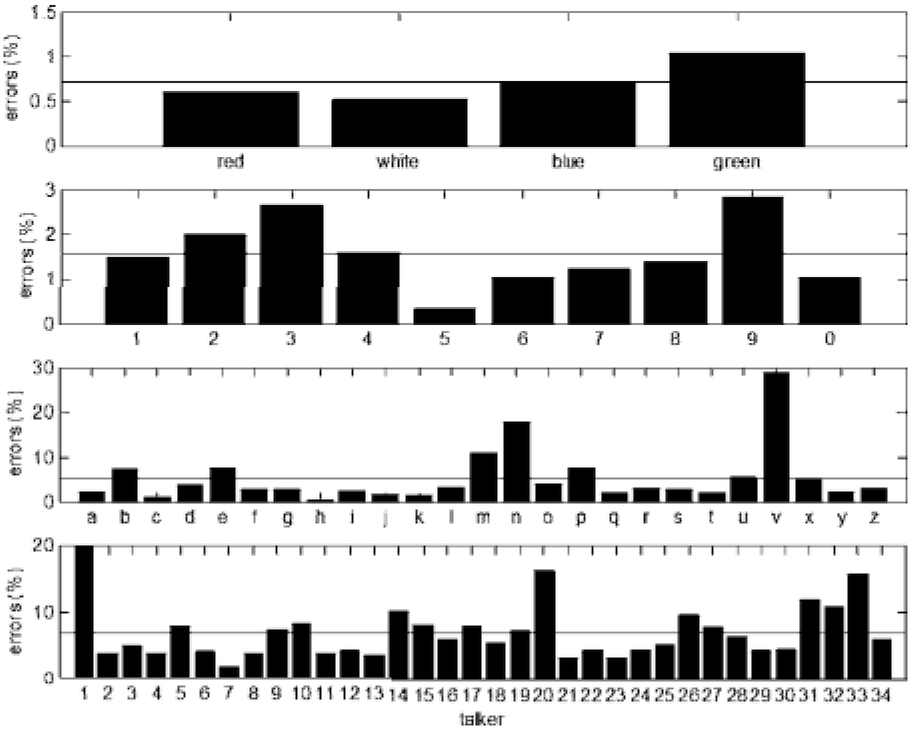

Figure 2: The percentage of utterances in which at least one keyword was incorrect is how talker error rates are calculated.

3. CONCLUSION

To support integrated computational-behavioral research in speech perception, a large multitalker audio-visual sentence corpus called Grid has been assembled. The spoken content is readily recognized in calm and low-noise circumstances, according to audio-only intelligibility tests. There will be more visual and audiovisual intelligibility tests in the future. The text is believed to have been recognized and found in word recognition, and the rectangular bounding box containing the text is provided. It is necessary to identify the word in the enclosing box. The techniques for doing word recognition may be divided into two categories: top-down and bottom-up approaches. A collection of terms from a dictionary is utilized in top-down methods to determine which word best fits the supplied picture. In majority of these techniques, images are not segmented. As a result, the top-down method is also known as segmentation-free recognition. The picture is divided into numerous components in bottom-up methods, and the segmented image is then sent through a recognition engine. To recognize the text, an off-the-shelf optical character recognition (OCR) engine or a custom-trained one is utilized[10].

REFERENCES:

[1] K. M. Yorkston and D. R. Beukelman, “A comparison of techniques for measuring intelligibility of dysarthric speech,” *J. Commun. Disord.*, 1978.

[2] M. Cooke, J. Barker, S. Cunningham, and X. Shao, “An audio-visual corpus for speech perception and automatic speech recognition,” *J. Acoust. Soc. Am.*, 2006.

[3] J. M. Liss, S. Spitzer, J. N. Caviness, C. Adler, and B. Edwards, “Syllabic strength and lexical boundary decisions in the

perception of hypokinetic dysarthric speech," *J. Acoust. Soc. Am.*, 1998.

[4] D. Estival, S. Cassidy, F. Cox, and D. Burnham, "AusTalk: An audio-visual corpus of Australian English," in *Proceedings of the 9th International Conference on Language Resources and Evaluation, LREC 2014*, 2014.

[5] C. Sheard, R. D. Adams, and P. J. Davis, "Reliability and agreement of ratings of ataxic dysarthric speech samples with varying intelligibility," *J. Speech Hear. Res.*, 1991.

[6] M. Nirupa, P. Prema, S. Vidhya, and T. Lazar Mathew, "Formant modification to improve intelligibility of dysarthric speech," *Int. J. Med. Eng. Inform.*, 2011.

[7] S. Selva Nidhyananthan, R. Shantha Selva kumari, and V. Shenbagalakshmi, "Assessment of dysarthric speech using Elman back propagation network (recurrent network) for speech recognition," *Int. J. Speech Technol.*, 2016.

[8] P. Upadhyaya, O. Farooq, M. R. Abidi, and P. Varshney, "Comparative Study of Visual Feature for Bimodal Hindi Speech Recognition," *Arch. Acoust.*, 2015.

[9] H. Kim *et al.*, "Dysarthric speech database for universal access research," in *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, 2008.

[10] O. Abdo, S. Abdou, and M. Fashal, "Building audio-visual phonetically annotated Arabic corpus for expressive text to speech," in *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, 2017.