A speech corpus for multitalker communications research

Robert S. Bolia, W. Todd Nelson, and Mark A. Ericson
Air Force Research Laboratory, Wright-Patterson Air Force Base, Ohio 45433

Brian D. Simpson
Department of Psychology, Wright State University, Dayton, Ohio 45435

(Received 1 October 1999; revised 18 October 1999; accepted 19 October 1999)

A database of speech samples from eight different talkers has been collected for use in multitalker communications research. Descriptions of the nature of the corpus, the data collection methodology, and the means for obtaining copies of the database are presented. © *2000 Acoustical Society of America.* [S0001-4966(00)00902-4]

PACS numbers: 43.66.Pn, 43.72.Kb [JMH]

INTRODUCTION

Several recent experiments at the Air Force Research Laboratory have investigated the utility of spatial audio displays for augmenting speech intelligibility in multitalker communications environments (Bolia *et al.*, 1999; Nelson *et al.*, 1998a; Nelson *et al.*, 1998b; Simpson *et al.*, 1999). Some of the goals of this research included: (1) an empirical determination of the maximal number of channels for which the benefits of spatialization may be realized in a spatial audio display designed to aid in the segregation of simultaneous, context-independent speech sources; (2) an evaluation of the efficacy of four different spatialization schemes for this task; and (3) the manner in which these factors interact with the sex of the target talker. In order to accomplish these goals, a large number of speech samples from talkers of both sexes were required. The purpose of this article is to describe the methods employed in collecting these speech samples, as well as the form of the resulting corpus, with the intent that other researchers in the field might benefit from their availability.

I. COORDINATE RESPONSE MEASURE (CRM)

The task selected for the sequence of investigations cited above was a version of the Coordinate Response Measure (CRM), a nonstandardized communication performance task adapted from similar tasks by Moore (1981) as a measure of speech intelligibility more relevant to military environments than standardized tests such as the Modified Rhyme Test. The phrases in the CRM consist of a call sign and a color–number combination, all embedded within a carrier phrase. Hence a typical sentence would be ‘‘Ready baron, go the blue five now,’’ where baron is the call sign, and blue five is the color–number combination. In the performance of the task, each listener is assigned a call sign, and responds by indicating the color–number combination spoken by the talker who uttered his or her call sign. If the listener does not hear his or her call sign spoken, he or she does not respond (or, equivalently, reports the absence of his or her target call sign). Possible dependent measures thus include the percentage of correct call sign detections and the percentage of correctly identified color–number combinations, as well as their associated reaction times.

The nature of the phrases in the CRM suggests its utility for measuring speech intelligibility in multichannel communications environments. McKinley *et al.* (1994) and Ericson and McKinley (1997) have employed it for such a purpose, and other researchers have devised and used similar tests (Koehnke *et al.*, 1998). Given a collection of multiple talkers speaking simultaneously, each speaking a different call sign and a different color–number combination, it is possible to interpret the percentage of correct call sign detections as a measure of the effectiveness of masking by competing speech signals, since, in order to make a correct detection, a listener must be able to distinguish his or her call sign from a collection of simultaneously spoken call signs. One can then interpret the percentage of correctly identified color–number combinations as a measure of a listener's ability to selectively attend to a single channel while ignoring irrelevant channels, since a listener must attend to a particular talker/location (i.e., the talker/location from which his or her call sign was spoken) in order to correctly identify the color–number combination emanating from that talker/location. In addition, it may also be desirable to perform a signal detection analysis using the detection portion of the task, from which can be gained measures of sensitivity (d') and response bias (β and/or c). The relatively context-free nature of the phrases ensures that changes in speech intelligibility are due to specific experimental manipulations rather than to contextual clues found in natural discourse.

Previous studies employing the CRM, such as those conducted by Ericson and his colleagues (Ericson and McKinley, 1997; McKinley and Ericson, 1997; McKinley *et al.*, 1994), have used live talkers as speech stimuli. Due to the large number of talkers required, as well as the need for precise control of stimulus onset, it was determined that digital recordings of talkers would be preferable to live talkers for the series of investigations cited above (Bolia *et al.*, 1999; Nelson *et al.*, 1998a, 1998b; Simpson *et al.*, 1999). It was for this reason that the speech corpus described herein was collected.

Factorial combinations of eight call signs (‘‘arrow,’’ ‘‘baron,’’ ‘‘charlie,’’ ‘‘eagle,’’ ‘‘hopper,’’ ‘‘laker,’’ ‘‘ringo,’’ ‘‘tiger’’), four colors (‘‘blue,’’ ‘‘green,’’ ‘‘red,’’ ‘‘white’’), and the numbers between one and eight yielded a total of 256 phrases, all of which were recorded by each of the eight talkers, for a total of 2048 phrases. Four males and four females, between the ages of 18 and 26, participated as talkers. None of the talkers had any reported or readily detectable speech pathology.

II. METHODS OF COLLECTION

The recording of the phrases was conducted in a sound-attenuated room of dimensions 3.93×3.40×3.50 m, normally used for conducting audiometric examinations. The walls and ceiling of the room were lined with sound-absorbing foam to reduce reflections; the floor was covered with commercial carpet. During the recordings, the talker was seated in an immobile office-type chair in one corner of the room. A Bruel & Kjaer Type 4165 1/2-inch microphone connected to a Bruel & Kjaer 2639 preamplifier was placed in a microphone stand approximately 3 cm in front of the talker. The output of the microphone was amplified using a Bruel & Kjaer Type 5935 dual microphone supply, and the resulting waveform digitized at the sample rate of 40 kHz by means of a Tucker-Davis Technologies (TDT) DD1 combined analog-to-digital/digital-to-analog converter. The recorded phrase was then converted back to an analog signal using the same TDT DD1, amplified via a Crown D-75 amplifier, and presented to the experimenter over Sennheiser HD-560 headphones.

The process of recording the phrases was controlled by a computer program. The talker was presented with a visual display containing the phrase he/she was to speak. Recording began when the talker pressed a key on the computer keyboard, and continued for a period of 3 sec. If the recorded phrase appeared to be ''correct,'' the experimenter accepted it and proceeded to the next phrase; if it were incomplete or spoken at an inappropriate pace, the experimenter rejected it and recorded it anew. Speaking rate was regulated by having the talker listen to a ''standard'' phrase—previously recorded by one of the experimenters—immediately prior to each recording, and pace himself/herself according to the pace of the standard. The latter practice ensured that, across talkers, similar phrases were of similar duration.

Once the corpus had been recorded in its entirety, all incipient silence was removed from each waveform for the purpose of synchronized playback. This and all subsequently described manipulations were accomplished with Cool Edit, a commercially available software product for the analysis and processing of acoustic waveforms. Following the synchronization of phrase onsets, each of the speech signals was bandpass filtered, with a passband extending from 80 Hz to 8 kHz. All of the phrases in the set were then scaled to have the same root-mean squared average power. Additionally, a second set of files was created from the original recordings in which the initial word ''ready'' was removed, so that the phrases could be synchronized at the onset of the call sign instead of the carrier phrase. All other manipulations to these phrases were identical to those in the first set.

III. AVAILABILITY OF CORPUS

The speech corpus described herein is available free of charge to researchers who send a blank recordable compact disk with a self-addressed postage-paid return envelope to the first author at the following address: Robert S. Bolia Air Force Research Laboratory (AFRL/HECP) 2255 H. St. Wright-Patterson Air Force Base, OH 45433-7022.

ACKNOWLEDGMENTS

The authors would like to acknowledge Michael L. Ward, of Veridian, and Dennis L. Allen, of Sytronics, Inc., for their contributions to the collection and modification of the speech samples; Raymond Slyh, of the Air Force Research Laboratory, for his helpful insights; Sue Vadnais and Anitha Reddy, both of whom participated in the collection of the speech samples as part of a summer internship program sponsored by Alliance for Education; and James Hillenbrand and two anonymous reviewers, for their helpful comments on an earlier draft of this article. A portion of this research was supported by the Air Force Office of Scientific Research.

Bolia, R. S., Ericson, M. A., Nelson, W. T., McKinley, R. L., and Simpson, B. D. (**1999**). ''A cocktail party effect in the mediam plane?,'' J. Acoust. Soc. Am. **105**, 1390–1391.

Ericson, M. A., and McKinley, R. L. (**1997**). ''The intelligibility of multiple talkers separated spatiallly in noise,'' in *Binaural and Spatial Hearing in Real and Virtual Environments*, edited by R. H. Gilkey and T. R. Anderson (Lawrence Erlbaum Associates, Mahwah, NJ), pp. 701–724.

Koehnke, J., Besing, J., Abouchaera, K. S., and Tran, T. V. (**1998**). ''Speech recognition for known and unknown target message locations,'' *Abstracts of the Twenty-First Midwinter Meeting of the Association for Research in Otolaryngology* (ARO, Mt. Royal, NJ), p. 105.

McKinley, R. L., and Ericson, M. A. (**1997**). ''Flight demonstration of a 3-D auditory display,'' in *Binaural and Spatial Hearing in Real and Virtual Environments*, edited by R. H. Gilkey and T. R. Anderson (Lawrence Erlbaum Associates, Mahwah, NJ), pp. 683–699.

McKinley, R. L., Ericson, M. A., and D'Angelo, W. R. (**1994**), ''3-dimensional auditory displays: Development, applications, and performance,'' Aviat., Space Environ. Med. **65**, 31–38.

Moore, T. J. (**1981**). ''Voice communication jamming research,'' *AGARD Conference Proceedings 311: Aural Communication in Aviation* (AGARD, Neuilly-Sur-Seine, France), pp. 2:1–2:6.

Nelson, W. T., Bolia, R. S., Ericson, M. A., and McKinley, R. L. (**1998a**). ''Monitoring the simultaneous presentation of multiple spatialized speech signals in the free field,'' Proceedings of the 16th International Congress on Acoustics and the 135th Meeting of the Acoustical Society of America, 2341-2342.

Nelson, W. T., Bolia, R. S., Ericson, M. A., and McKinley, R. L. (**1998b**). ''Monitoring the simultaneous presentation of spatialized speech signals in a virtual environment,'' Proceedings of the 1998 IMAGE Conference, pp. 159–166.

Simpson, B. D., Bolia, R. S., Ericson, M. A., and McKinley, R. L. (**1999**) ''The effect of sentence onset asynchrony on call sign detection and message intelligibility in a simulated 'cocktail party','' J. Acoust. Soc. Am. **105**, 1024.