Proceedings of the Institute of Acoustics

DESIGN, COLLECTION AND ANALYSIS OF A MULTI-SIMULTANEOUS-SPEAKER CORPUS

Malcolm Crawford, Guy J. Brown, Martin Cooke and Phil Green

Speech and Hearing Research Group, Department of Computer Science, University of Sheffield, Regent Court, 211 Portobello Street, Sheffield S1 4DP, England.

{m.crawford, g.brown, m.cooke, p.green}@dcs.shef.ac.uk

1. ShATR: A CORPUS OF AUDITORY SCENES

Spoken communication usually takes place in an acoustically cluttered environment — there are typically several sound sources present, whose number and characteristics cannot be pre-determined. The *Sheffield-ATR*[1] multiple-simultaneous speaker database — ShATR — is a new corpus which has been collected (although not yet fully annotated or released) to facilitate research on speech perception in such natural surroundings. The purpose of this paper is to describe the rationale for, collection of, and immediate plans for this data.

Human listeners have a remarkable ability to separate out and pay selective attention to individual sound sources, a feat referred to as "auditory scene analysis" (ASA; Bregman, 1990). Work at Sheffield (Cooke 1993, Brown & Cooke, 1994) has achieved some success in computational modelling of ASA based on primitive, bottom-up grouping principles such as common onset, periodicity and good continuation. We have also begun to address the problem of how ASA might be used in the task of automatic speech recognition (ASR) in noise (Cooke, Green & Crawford, 1994).

Computational ASA research has now reached the point where a corpus of auditory scenes is required for training and evaluation of segregation and recognition algorithms. Most existing speech corpora (e.g. TIMIT, the Resource Management and Wall Street Journal corpora), however, consist of clean speech. The NOISEX database provides speech with added noise, of various types and at various SNRs: this material, however, is not typical of auditory scenes; e.g. there is only one noise source whose characteristics do not change and it is added to clean speech, so there is no Lombard effect (*cf.* Summers *et al.*, 1988)[2].

We set out to record a multiple sound-source corpus with the following requirements:

(a) There are several speakers;
(b) The speakers are engaged in a collaborative task;
(c) The task is sufficiently natural to provoke spontaneous speech;
(d) The execution of the task generates additional non-speech noises;
(e) For a significant proportion of the time more than one sound source is active;
(f) Speech material is generated which will facilitate both large-vocabulary speech recognition and common keyword-spotting.

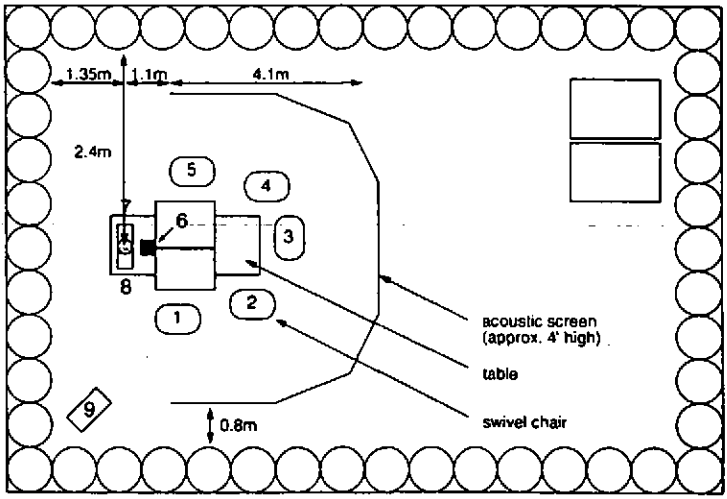
2. THE CROSSWORD TASK

The task we devised to meet these requirements was based on solving crossword puzzles. The task has an unrestricted vocabulary, but induces frequent occurrences of a few words — "across", "down", "blank", the

1. Advanced Telephony Research, Kyoto, Japan.
2. Texas Instruments' CCDATA corpus (*cf.* Hansen, Womack & Arslan, 1994) does contain speech spoken in noise, but from two-way voice communications across handsets — an impoverished auditory scene.

numerals from 1 to 30, and so on — assuring (f). Pilot studies led us to adopt the following arrangement:

- There were five speakers — see requirement (a), seated around a table;
- There were two teams each of two speakers, each team solving a different crossword puzzle (b);
- The fifth speaker had the answers to both crosswords, and was allowed to give hints;
- Seating was arranged so that it was necessary for each team to "talk across" the other team, and to the hint-giver. A diagram showing the layout of the recording environment is given in figure 1, with a photograph in figure 2. The hint-giver took position 3, and participants (1, 4) and (2, 5) paired into teams. This served to create a more interesting acoustic environment than (1, 2) and (4, 5) pairings since more speech was directed across the mannikin as there was no "huddling".



channel	component	equipment type
1 - 5	participant, wearing headset microphone	RAMSA WM-S10
6	omnidirectional (pressure zone) microphone	Crown PZM30
7,8	mannikin	B&K type 4128, with B&K ears, type 4158 (R) and 4159 (L); microphones type 4134 powered by a B&K power source type 2804.
9	video camera	ceiling-mounted

FIGURE 1. Plan of the recording chamber and equipment set-up (not to scale).

DESIGN, COLLECTION AND ANALYSIS OF A MULTI-SIMULTANEOUS-SPEAKER CORPUS

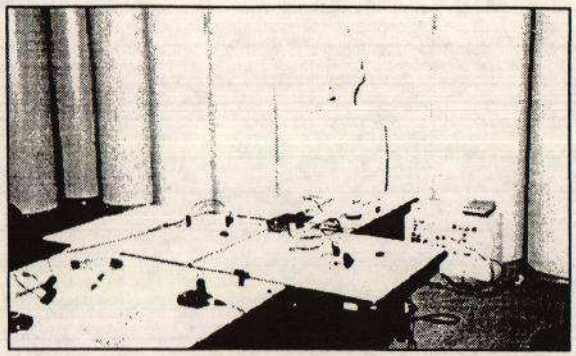


FIGURE 2. Photograph of the recording chamber.

We found that the task was sufficiently absorbing, and sufficiently collaborative (semi-collaborative, one might say) to fulfil (c) above, and that there were abundant non-speech noises such as moving chairs, pens writing and so on to satisfy (d). We developed an algorithm to measure the degree of overlap and confirmed that (e) was fulfilled.

3. DATA COLLECTION

Recording equipment

At total of eight microphones were used (*cf.* figure 1), whose signals were fed to a Yamaha HA-8 8-channel microphone preamplifier and then to a TASCAM DA-88 digital 8-track recorder, located outside of the recording chamber. Signals were recorded at 48kHz[1], 16-bit quantisation. A plot showing the waveforms for the eight channels side-by-side for a one-minute extract from a session is given in figure 3. In addition, a video recording was made of the sessions from a ceiling-mounted camera.

Room acoustics and equipment calibration

Recordings were made in the variable reverberation chamber at ATR, Kyoto. The room was configured to give an average reverberation time of around 0.35s. In order to more accurately measure room acoustic qualities, reverberation time and impulse response, a number of "calibration" recordings were made, with the loudspeaker placed at each talker location in turn:

M-sequence (*cf.* Golomb, 1964), two repetitions of 10-cycle-sequences;

Ten clicks (single pulse of an 8kHz square wave);

Ten time-stretched pulses (Aoshima, 1981).

Sounds were delivered through a Studio Monitor 4410 loudspeaker, supplied by an Accuphase P800 amplifier. Recordings were made from the mannikin ear microphones, the omnidirectional microphone, and a directional microphone (B&K type 4134) placed directly in front of the speaker.

In order to calculate the impulse response of the recording and delivery equipment itself, a second series of recordings was made using the same stimuli, but in an anechoic chamber. Additionally, a recording was made

1. 48 kHz was chosen as an emerging standard for professional quality recordings, and for ease of downsampling to rates more commonly used in ASR, e.g. 16 kHz.

DESIGN, COLLECTION AND ANALYSIS OF A MULTI-SIMULTANEOUS-SPEAKER CORPUS

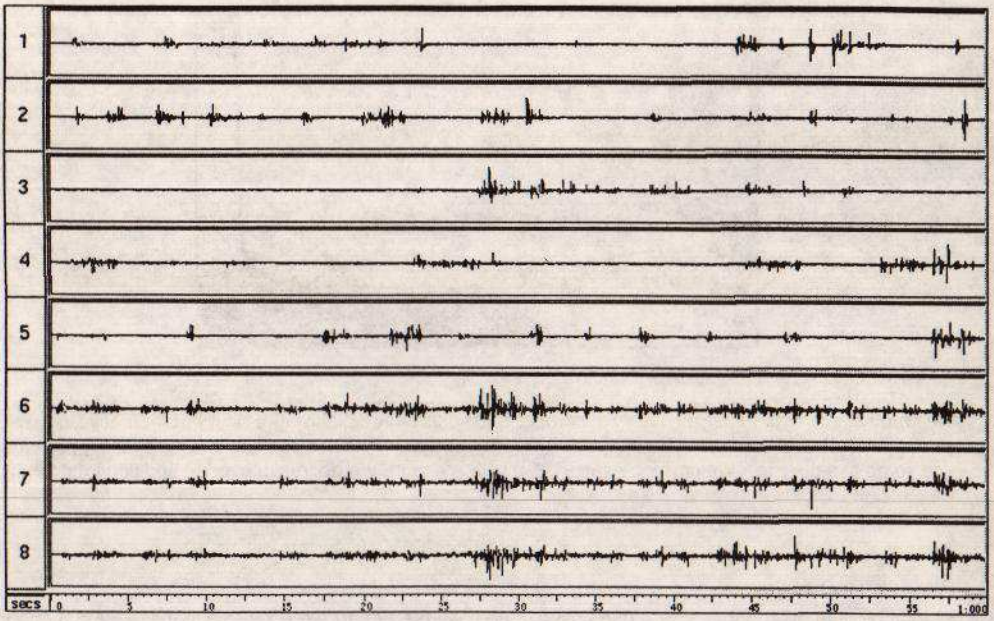


FIGURE 3. Plot showing waveforms of the eight channels recorded for a one-minute segment extracted from the database: refer to figure 1 for details of each channel. It is clear that at various times more than one participant is talking simultaneously.

from each ear of a 123.5dB calibrated input from a B&K type 4220 pistonphone; this will allow estimates of SPL to be calculated for other input.

Enrolment

Prior to taking part in the task, each subject went through an enrolment procedure to provide data which could be used for training, or adapting, ASR systems, or speaker identification systems. This consisted of reading the following whilst sitting the same location in which the speaker would be during the task:

The TIMIT shibboleth sentences, twice each;

Ten repetitions each of the words "across", "down", "yes", and "no", interspersed in random order;

Ten repetitions each of the letters of the alphabet and digits from 0 to 30 (in random order);

A passage from The Japan Times.

Recording sessions

A total of four sessions were recorded, each lasting around 30 minutes. In the first three, all participants remained seated, whilst in the fourth the hint-giver was permitted to walk around the central table in order to provide a moving noise-source, and two other speakers entered the room about half-way through the session.

DESIGN, COLLECTION AND ANALYSIS OF A MULTI-SIMULTANEOUS-SPEAKER CORPUS

4. DATA ANALYSIS AND SUPPORTING TOOLS

The utility of speech corpora is heavily dependent on the quality of the supporting analysis. Whilst linguistic analysis has — not surprisingly — been at the forefront of many corpora, past collections have typically failed to provide software tools to allow researchers to access the material in a flexible, task-dependent fashion on a variety of platforms, in a host of common formats.

Our intention is to transcribe each of the individual speaker signals at the word-level, and to use an automatic phone alignment technique to produce a more-detailed transcription. In the initial release, the automatic transcription will not be manually corrected — its main role will be to allow for rapid access to known sound classes. Additionally, we will transcribe non-speech material using a shallow taxonomy (there are a limited number of non-speech sources present in the collected material, but they occur quite frequently)[1].

We intend to provide software tools to allow easy access to specified segments. Since we have multiple speakers, there is an additional dimension to any such queries. We wish to handle arbitrary queries, to include:

"Find all vocalic segments which overlap with at least one other talker";

"Find all overlaps between talker 2 and talker 5 of at least 1s duration";

"Find all transitions between talker 2 and talker 3".

We will supply omnidirectional and binaural channels at the full 48 kHz sampling frequency, but will downsample the remaining 5 channels to 16 kHz. Laboratories may obtain the full 48 kHz individual microphone signals if required. Data will be in 16-bit NeXT .snd (Sun .au) format, but we will provide routines to access the material from a variety of platforms, and to convert it to other common formats.

At present (prior to detailed analysis of the sessions), we expect to release a single full session (about 30 minutes in duration). Without compression, this amounts to just over 0.8 GByte of material, excluding transcriptions, calibration results and enrolment data. The latter is around 0.32 GByte (5 speakers, 20 minutes each at 16 kHz). We will also make available copies of the video-recording to those who request it. At the time of going to press, the corpus is barely downloaded. However, we expect to have completed the analysis by the end of 1994, with software tools following in the first quarter of 1995.

We further intend to make full use of Internet resources to allow individuals to download small (but usable) quantities of material from the corpus[2]. We will also specify these "usable" segments as test material for comparative evaluations of ASA systems, psychoacoustic studies etc. Additionally, we will set up and maintain an electronic forum for discussion of the corpus, and attempt to maintain a bibliography of all published works which make use of the material.

5. WHO WILL USE THIS CORPUS?

The primary purpose of the ShATR corpus is to allow the performance of algorithms for segregating simultaneous sound sources to be empirically evaluated and compared on standard test data using standard metrics. Furthermore, since the corpus contains mono and stereo recordings, it can be used to assess the relative performance of monaural and binaural segregation schemes. The data collected in this corpus will be also useful, however, in a number of other domains.

---

1. *Cf.* also our hopes for future developments in the final section.
2. From January 1, 1995, details will be available through pages on the World-Wide Web server at `http://www.dcs.shef.ac.uk`.

Localisation studies

The binaural recordings will provide data for localisation studies. A number of systems have been proposed for segregating sounds according to their spatial locations, but typically these have used simple test stimuli, such as the simultaneous speech of two talkers with different average pitch ranges (e.g., Denbigh & Zhao, 1992). The ShATR corpus will provide test conditions which are much closer to those encountered in natural listening environments. Related to this point, there has recently been some criticism of Cherry's (1954) assertion that spatial location is a powerful cue for segregating voices in a "cocktail party" situation (Meddis, 1994). The corpus of binaural recordings described here will allow an experimental assessment of whether the separation of sound sources in space leads to increased intelligibility of the target source as well as increased signal-to-noise ratio[1].

Automatic speech recognition and speaker identification

Increasingly, research in ASR is becoming focused on the problem of recognition in adverse environments (i.e. outside an anechoic chamber). The environment in which our corpus was recorded was reverberant and contained the voices of several talkers; it therefore qualifies as "adverse". Certainly, the corpus contains sufficient training data for word spotting studies. Additionally, an interesting possibility is that information about spatial location could be used to isolate the voices of single speakers and hence "bootstrap" a recognition algorithm. In turn, the partially trained recogniser could provide top-down information for the spatial segregation algorithm, so that segregation and recognition proceed in a symbiotic manner. Speaker identification algorithms could be implemented in a similar way.

Other applications

We anticipate that the corpus will find other applications in speech and language research. For example, a prosodic analysis of the corpus would be interesting in its own right, and could also provide the basis for intentional, dialogue and discourse analyses. A further possibility is that the video recordings could be used to study the role of visual cues in speech perception within a multi-speaker environment.

6. FUTURE CORPORA

One observation made during the recording sessions was that speakers rarely faced the manikin, presumably because it did not actively participate in the conversation. Future recordings could address this issue by allowing a remote participant to listen in real time to the binaural recordings, responding interactively through a speaker placed in the mouth of the manikin. This scenario could be extended to the visual domain by equipping the manikin with stereo cameras, and making close-up video recordings of the participant's lip movements. Such an arrangement might provide a powerful means of studying multi-modal interaction between speakers and listeners, and would also allow the study of speech perception with and without lip-reading cues in a realistic multi-speaker environment. Finally, future recordings could dispense with the manikin completely and make recordings from in-ear microphones fitted to the participants.

It was also noted that voice quality during the final recordings varied less than during practice sessions — participants were concerned to remain "serious", and avoid any utterances which it might be unwise to release on a CD-ROM. Future recordings might seek to promote still more natural and lively exchanges, with greater variation in voice quality.

Finally, we hope that other researchers using this data will make their own analyses available for inclusion on subsequent update releases of this database. This will ensure a growing utility of the resource, and further facilitate comparison of analyses and methodologies employed by various groups[2].

1. It should be noted, however, that the lack of realism of head movement at the binaural receptors may adversely affect the assessment.

ACKNOWLEDGMENTS

This work was supported by SERC Image Interpretation Initiative Research Grant GR/H53174; a study visit grant from Advanced Telephony Research, Kyoto, to Guy Brown and Malcolm Crawford; Royal Society and Royal Academy of Engineering awards to Phil Green, and a Royal Society grant to Martin Cooke.

The authors express their gratitude to Minoru Tsuzaki, Hideki Kawahara, Hiroata Kato, and Masako Tanaka of ATR for their invaluable assistance and advice, to David Kirby of the BBC for much help and support during the early stages of configuring the recording equipment, and to Inge-Marie Eigsti of ATR for her willing participation as a subject.

We also acknowledge help, advice and criticism from many colleagues in the speech, hearing and linguistics communities.

REFERENCES

N. Aoshima (1981), "Computer-generated pulse signal applied for sound measurement", *JASA*, **9**, 1484-1488.

A.S. Bregman (1990), *Auditory Scene Analysis*, MIT Press.

G.J. Brown & M.P. Cooke (in press), "Computational auditory scene analysis", *Computer Speech & Language*.

E.C. Cherry (1954), "Some experiments on the recognition of speech with one and with two ears", *JASA*, **25**, 975-979.

M.P. Cooke (1993), *Modelling Auditory Processing and Organisation*, Cambridge University Press.

M.P. Cooke, S.W. Beet & M.D. Crawford (eds) (1993), *Visual representations of speech signals*, Wiley, ISBN 0 471 93537.

M.P. Cooke, P.D. Green & M.D. Crawford (1994), "Handling missing data in speech recognition", *International Conference on Speech and Language Processing*, Yokohama.

M.P. Cooke & G.J. Brown (1994), "Separating simultaneous sound sources: issues, challenges and models"; In: *Fundamentals of Speech Synthesis and Speech Recognition*, E. Keller (ed.), John Wiley & Sons, 295-312.

M.P. Cooke & G.J. Brown (1993), "Computational auditory scene analysis: Exploiting principles of perceived continuity", *Speech Communication*, **13**, 391-399.

P.N. Denbigh & J. Zhao (1992), "Pitch extraction and separation of overlapping speech", *Speech Communication*, **11**, 119-125.

W.S. Golomb (1964), *Digital Communication*, Prentice Hall.

J.H.L. Hansen, B.D. Womack & L.M. Arslan (1994), "A source generator based production model for environmental robustness in speech recognition", *International Conference on Speech and Language Processing*, Yokohama.

R. Meddis (1994), "Human auditory signal processing of binaural signals", *Proceedings of the ATR Workshop on A Biological Framework for Speech Perception and Production*, Kyoto, September 16-17.

W. Summers *et al.* (1988), "Effects of noise on speech production: Acoustic and perceptual analyses", *JASA*, **84** (3), 917-928.

2. *Cf.* also the goals set out in Cooke, Beet & Crawford (1993).