Proc. 1996 Australian New Zealand Conf. on Intelligent Information Systems, 18-20 November 1996, Adelaide, Australia. Editors, Narasimhan and Jain

Cocktails and Brainwaves
Experiments with Complex and Subliminal Auditory Stimuli

David M. W. Powers, Ph.D, Director,
Simon E. Dixon, Ph.D.,
and
C. Richard Clark, Ph.D, Director,
Darren L. Weber, B.Sc.(Hons)

Artificial Intelligence Laboratory.
Department of Computer Science
The Flinders University of South Australia

Cognitive Neuroscience Laboratory
School of Psychology
The Flinders University of South Australia

Email: {David.Powers,Richard.Clark,Simon.Dixon,Darren.Weber}@flinders.edu.au
Phone/Fax: +61-8-201-{3663,2425,3664,3580} / +61-8-201-{3626,3877}
Change: **Adelaide numbers will change in August when our prefix becomes +61-8-8201**

This paper deals with the problem of processing acoustic signals originating from multiple sources in a potentially noisy environment. Previous research in speech processing and cognitive modelling has tended to concentrate on single sources and relatively noise-free signals. Separating out different signals from a multitude of sources is a significant part of human auditory processing. In speech processing research, the problem we are dealing with is known as the cocktail party syndrome. The processing of polyphonic music involves similar challenges, and auditory scene analysis (ASA) has been proposed as a means of separating out component signals and identifying their sources. In subliminal auditory processing, a speech signal which is masked from conscious awareness by a music signal provides an extreme form of the multiple source problem and permits exploration of the boundary between conscious and unconscious auditory processing. The research presented employs machine learning and associative models to characterize and track individual signals, and uses electroencephalographic (EEG) analysis to more precisely characterize human processing of multimodal signals.

Keywords: Auditory Scene Analysis; Speech Processing; Automated Transcription; Machine Learning; EEG

INTRODUCTION

In this paper we focus on the problem of dealing with acoustic signals originating from multiple sources in a potentially noisy environment. Despite significant progress in developing systems and models for limited natural language and speech processing, this area remains a challenge to artificial intelligence researchers, cognitive scientists and electrical engineers. In particular, speech processing traditionally assumes limitations such as customization for a single speaker, close-miking, neglible ambient noise, and often other conditions including the requirement for words to be enunciated separately. Similarly, when processing acoustic music signals, it is difficult to reconstruct the score of a polyphonic composition.

Recently, interest in a more integrated cognitive science approach has led to models in which multiple sources of information are used to attack perceptual and cognitive processing problems, and auto-correlative, connectionist, statistical and learning-based techniques are employed to analyze linguistic input [7,8]. Furthermore, there is evidence that associative processing is employed to discover characteristics of a signal and selectively filter out other signals, and that spectral similarity may be used to mark internal neural signals as relating to the same event or concept [1]

Developments in the area of multi-electrode EEG permit the topographic mapping of scalp electrical activity that is volume conducted from underlying brain structures. These scalp potentials reflect the gross neural activity of localised regions of activity in the brain. With language processing, much of this activity is presumed to originate in the cortex. This is the outer layer or "mantle" of the brain and the part mainly responsible for fine-grained sensori-motor and associative processing [3]. While there is a general and long standing expectation that musical processing is handled by the 'right brain' (in non-musicians) and language processing by the 'left brain' (in a normal right-hander), there is scope for more detailed analysis of the areas of the cortex which are active for different kinds of auditory processing, and of the nature of the activity when multiple signals are present.

In addition, variation of the relative strength of two such signals in relation to the threshold of conscious perception and the masking effect, allows examination of which areas of the cortex are involved in conscious and unconscious processing. More detailed analysis should also be able to distinguish certain semantic and syntactic aspects of the message as well as distinct characteristic frequency spectra associated with the different signals [3,6].

AIMS AND OBJECTIVES

Engineering Objectives

Our approach to this problem area is proceeding on a number of fronts, and this paper will outline the subprojects and report some preliminary results. Our primary objective is to apply associative, statistical and learning techniques to the multisource auditory problem in order to increase the effective signal to noise ratio for an attended source and to enhance processing of a source without necessarily separating it out explicitly. Cleaning up signals is an application in its own right, but our aim is simply to transcribe or understand one or more of the signals.

In order to tackle any problem, it is important to have an appropriate representation of the input data. In the case

of auditory data, this should take into account the known characteristics of the human ear, and should retain information which is useful whilst discarding what is not. The decision as to what constitutes useful information depends, however, on our purpose. If our aim is to identify a speaker or an instrument we would tend to retain information about spectral character which is typically discarded by systems that seek simply to understand speech or transcribe a melody.

For our purposes, the character of a voice (whether human or instrumental) is information which we want to retain and use to distinguish our target voice from other noise and voices. Conversely, traditional large vocabulary speech systems are tuned to an individual voice using an extensive training process. We want to build a model which is economical, rather than providing a brute force dictionary of typical speech units at some level. We will then use this model not just to switch the system to a particular speaker, but to be able to track our target speaker in the face of background noise and voices. Finally, the tuning in to a speaker should not involve a studio quality training phase, but should be possible *in situ*.

The technology which we are employing for our linguistic analysis involves creating statistical models and autocorrelating the signal — that is we find parts of the signal that are similar in *content* or *context* with other parts of the signal, and classify those segments into classes. In application to speech, previous results include associating of similar speech vectors and the identification of phonetic classes [8]. At higher levels, syllable structure, syntactic relations and semantic classes can be identified [7,8].

Scientific Objectives

Notwithstanding the many practical applications for the technology we are seeking to develop, this project sees the pure scientific objectives of understanding and modeling auditory processing as being of considerable importance in its own right, and thus a worthwhile aim for the project. The problem area is therefore being explored as a multifaceted interdisciplinary research project in Cognitive Science. We believe that scientific understanding is a prerequisite for technological achievement and that there should be a symbiotic relationship between theoretic models and practical systems in this area.

In order to increase our understanding of human processing of multimodal auditory signals, we are seeking to develop the capacity to make verifiable predictions from our proposed models/systems and to confirm these predictions using the methods of cognitive electrophysiology.

These predictions may take many forms. The analytic techniques which we are using result in models which show certain structures and sequences of emergence which do not correspond to standard Linguistic and Psycholinguistic expectations. For example, the role of closed class words, functional affixes and prosody is recognized at the earliest stages of processing whereas the tacit assumption has been that the late emergence of functional features in the speech of young children means that they do not have any significance to them and are not recognized. Different areas show up in the EEG maps for types of words, but we also need a mechanism for distinguishing areas involved in conscious and preconscious processing, and for this purpose we are using subliminal auditory conditions.

SYSTEMS DEVELOPMENT

Fourier Preprocessing

Although the techniques used in this project are based on Fourier analysis, it is recognised that this method cannot provide a sufficiently high resolution in both the time and frequency domains simultaneously [4]. Nor is there an optimal compromise for all recognition tasks, as different domains require the identification of different acoustic features. The parameters chosen for a Fourier transform provide a tradeoff between resolution in the frequency domain and in the time domain.

For example, the recognition of speech requires high time resolution (tens of milliseconds), but a relatively low frequency resolution (50 or 100 Hertz is sufficient), whereas Western music can be recognised with a time resolution closer to 100ms, but requires a frequency resolution of less than a semitone, which is around 2 Hz at the lowest frequencies of interest. Suitable setting of parameters comes close to meeting the resolution needs in either domain, but the combined requirements are not directly attainable.

To counteract this problem, sounds are analysed at multiple resolution levels, and features are linked between the levels. For example, in a music recognition system, it appears to be sufficient to work at two levels, one for detecting frequency (pitch) accurately and one for detecting timing (rhythm) accurately. The resolution problem is thus solved at the expense of added complexity in grouping the acoustic data, but this approach has the additional virtue of consistency with the similar dual processing performed in the cochlea: the inner surface of the basilar membrane provides low frequency resolution whilst the outer surface permits high frequency discrimination.

Auditory Scene Analysis

In order to make sense of complex signals from multiple sources, a sound recognition system must be able to organise the components of sounds into groups which correspond to the acoustic sources. This appears to be done preconsciously in the human brain, and extensive psychological research in the area of auditory scene analysis (ASA) has identified a number of principles which reflect this area of brain functionality [1,2]. The grouping principles of ASA are common to other areas of basic cognitive function (such as vision), and also share common ground with Gestalt psychology.

In agreement with ASA principles, we are examining two stages of grouping: primitive and schema-based. Primitive grouping is used to sort raw elements of sensory data, whereas schema-based grouping builds world-level descriptions of acoustic sources and events from the output of the primitive grouping stage.

Primitive grouping itself can be divided further into two types: sequential or horizontal grouping, based on proximity of acoustic components, and simultaneous or

vertical grouping, based on similarity of components. The process of auditory scene analysis can be modelled as a collection of acoustic components, each initially assigned to its own group, competing to attract other components to merge groups together.

Horizontal grouping is determined by frequency proximity, temporal proximity, spectral overlap, spectral profile (e.g. formants), relative strength of harmonics, slope of frequency change and spatial cues. The nearer any two components are according to each of these criteria, the more strongly the two groups will compete to merge. Vertical grouping is based on frequency proximity, harmonicity, common fate (onset and offset synchrony, common amplitude or frequency modulation), and spatial correspondence. The greater the similarity of components with respect to these principles, the more likely they will be attributed to the same acoustic event.

Dynamic Modelling

The higher levels of auditory scene analysis are based on matching recognised patterns to previously learned schemas. That is, the brain matches sounds to expected patterns, which allows the filling in of gaps in the sensory data, and the correction of errors in earlier levels of processing.

Each of the ASA principles relates directly to expectations of acoustic events in the world, and enables the sorting of sound components into their most likely groups. In a computational setting, we view the groupings produced by these principles as supporting a particular model of the acoustic environment, and by building these models dynamically as the sounds are processed, we are able to develop expectations which which should themselves infuence the way the data is processed. This feedback allows the system to be fine-tuned without being hard-wired for any specific task or voice.

NEUROSCIENCE EXPERIMENTS

Subliminal Auditory Stimuli

The combination of musical and speech signals has been controlled by using techniques which are standard in relation to developing subliminal programming. Following the standard approach recommended in [10,11], we arranged for the local intensity of a speech signal to track the local intensity of a music signal and produced tapes of music and speech with various degrees of relative intensity, giving conditions with pure speech, pure music, music with supraliminal (above threshold) speech and music with subliminal (below threshold) speech. The pure music and pure speech conditions acted as control conditions for the purpose of statistical analysis. This provided the basis for identification of scalp regions with EEG activation related to supraliminal and/or subliminal speech processing. In addition, we varied between isolated word and continuous speech conditions and we controlled for the emotive content of the words. The music took the form of extracts from a single tape of relaxation music, all of which were very similar in character.

Although other researchers [10] have demonstrated detectable physiological responses to subliminal auditory stimuli, no previous experiments using EEG are known.

EEG Experiments and Results

The study tested two right-handed subjects (18-25 years) who had normal hearing. The subjects completed each of the auditory stimulus conditions described in the section above whilst EEG activity was recorded from 19 scalp sites according to the 10-20 convention [5]. EEG was sampled and held every 1 milliseconds using an analog-to- digital converter with 16 bit resolution, and amplified 1000 times using DC amplifiers (Neuroscan SYNAMPS). A linked ear reference was used. Electro-oculographic (EOG) activity was also recorded from the supraorbital ridge of one eye and referred to the outer canthus of the same eye.

EOG artefacts were removed from the digitised EEG using the algorithm of Semlitsch [9]. Any region of EEG containing electromyographic activity, or peak activity in excess of +/-100 microvolts, were also removed from the data prior to analysis. Data were then epoched individually into segments of continuous EEG of 2048 milliseconds duration. The data for each subject were then analysed using a Fast Fourier Transform (FFT) to identify EEG power at each electrode site for each stimulus condition in each of the following bands: alpha (8-13 Hertz), Beta 1 (14-20 Hertz), Beta 2 (21 - 34 Hertz) and Gamma (35-45 Hertz). A minimum of 121 EEG epochs was used for each FFT analysis.

Whilst this pilot experiment was undertaken with only two subjects, both demonstrated distinctive response to music and speech. T-tests have been used to assess the effect of subliminal and supraliminal speech on the EEG. For supraliminal speech this was achieved by comparing EEG spectral power for the supraliminal speech+music condition with the control music alone condition. Similarly, subliminal speech effects were assessed by comparing the subliminal speech+music condition with the music alone condition.

The effects for one subject are shown in Figure 1. Note that shaded regions show the areas of significant increase in power associated with subliminal and supraliminal speech processing ($p<0.01$, one-tail). Supraliminal speech caused left and right hemisphere

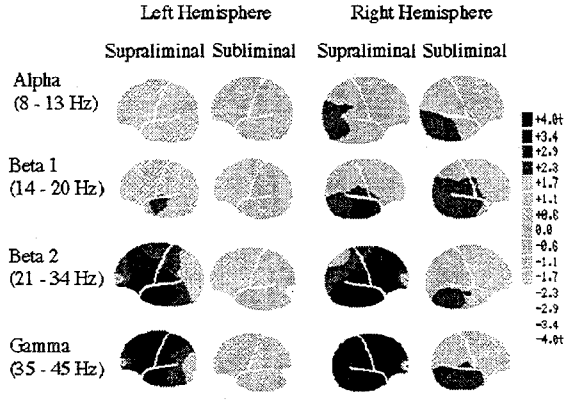

Figure 1: *The darker areas* (T-test>2.3) *identify regions of activation related to the addition of supraliminal and subliminal words to music* ($p<0.01$, one-tail).

activation in both the Beta 2 and Gamma bands with maximum power over the anterior temporal and posterior frontal regions. Right hemisphere activity was also evident in the posterior occipital regions for the Alpha band, but reached significance only for one subject. Left and right hemisphere activation covered a number of regions conventionally associated with auditory language processing. Subliminal speech failed to generate any left hemisphere activation but showed activation over regions of the right hemisphere and in common with a number of regions also affected by supraliminal speech. Several of these regions are also associated with auditory language processing. Overall, these results support the view that that subliminal speech is processed within the brain but indicate that this processing is constrained to regions in the right (rather than the left) cerebral hemisphere.

CONCLUSIONS AND FUTURE WORK

Our starting point is the assumption that all information in the signal should be used in our speech and music understanding task. In particular, the information which is used in identifying a speaker or an instrument, such as in speaker recognition systems, should be used to help track voices in a complex signal.

Our preliminary electroencephalographic experiments demonstrate that different characteristics of the different classes of signal, speech or music, are detected by the brain even when presented at levels below the threshold of conscious awareness. This suggests that the human auditory system does indeed make use of signal/source characteristics in processing complex signals. Together with previous research on auditory scene analysis this confirms that we should be looking for source characteristics, and is consistent with a source model view of speech processing.

More detailed experiments and analysis should allow (dis)confirmation of details of models being employed and will provide more information on the way the brain separates and processes auditory signals.

These initial experiments have also highlighted some problems with the experimental method which we will want to address in future EEG experiments. Indeed, we had hoped to obtain useful event related potentials (ERP) from these subjects — the ERP technique involves averaging over the events by class, with EEG epochs time-locked to the onset of each speech stimulus, but we found that the data was not of sufficient quality to allow this finer grained analysis.

The most apparent problem was that subjects tended to fall asleep during the experiments. Having relaxation music as the substrate for the experiments, this is not unexpected, but it is a factor which is difficult to control and is independent of the factors we do control. Preparing a subject for the experiment, including the correct seating of all the electrodes, is itself a long and somewhat stressful process. Although we detected a tendency for subjects to close their eyes during the calibration phase of the experiment, and requested that they try to keep them open, we found that the data was nonetheless contaminated by alpha waves characteristic of dozing (in the range 8-13Hz).

In this experiment we aimed to have subjects as receptive as possible for subliminal stimulation, and indeed have excellent evidence of subliminal effects. For this reason we tried to avoid distractions, and we avoided asking them to concentrate on anything. But this means that we have no control over what they are thinking or their state of consciousness. However to keep them awake it now seems appropriate to ask them to do some specific task which will occupy them while the signals are being presented.

We also noted that the intensity tracking arrangement recommended in [11] tended to make the supraliminal words less intelligible than we had hoped. Although this did not appear to affect comprehensibility of the continuous speech samples, it seems to introduce ambiguity which negates the utility of the samples based on random words (tagged for emotive content).

These experiments also serve to demonstrate that we have the capacity to construct and manipulate complex signals as well as to study their effects on EEG. This paper has not reported on computational experiments in analyzing speech and music signals, but initial experiments on separating sources are in progress.

ACKNOWLEDGEMENTS

Parts of this project were undertaken in association with the CSIRO Student Research Scheme and involved participation by students from South Australian High Schools as assistants and subjects.

REFERENCE

1. Bregman, A. (1990) **Auditory Scene Analysis: The Perceptual Organisation of Sound**, MIT Press
2. Brown, G and Cooke, M (1994) *Computational auditory scene analysis*, **Computer Speech and Language 8**:297-336
3. Clark, C.R., Pomeroy, D.E. and Tizard, J. (1995) *Neurocognitive pattern classification of distributed brain electrical activity.* In P. Slezak, T Caelli and C. R. Clark (eds) **Perspectives on Cognitive Science: Theories, Experiments and Foundations**. Ablex: New York
4. Dixon, S.E. (1996) *Multiphonic Note Identification*, **Australian Computer Science Communications 18**: 318-323
5. Jasper, H.H. (1958) *Report of the committee on methods of clinical examination in electro-encephalography*, **Electroencephalography and Clinical Neurophysiology 10**: 370-375
6. Neville, H.J. (1992) *Fractionating Language: Different Neural Subsystems with Different Sensitive Periods*, **Cerebral Cortex 2**: 244-258
7. Powers, D.M.W. (1996) *Unsupervised learning of linguistic structure: An empirical evaluation*, **International Journal of Corpus Linguistics 1**#2
8. Schifferdecker, G (1994) *Finding Structure in Language*, **Masters Thesis**, University of Karlsruhe FRG
9. Semlitsch, H.V., Anderer, P, Schuster, P. and Presslich, O. (1986) *A solution for reliable and valid reduction of ocular artefacts applied to the P300 ERP*, **Psychophysiology 23**: 695-703
10. Urban, M.J. (1992) *Auditory Subliminal Stimulation: a re-examination*, **Perceptual and Motor Skills 74**: 515-541
11. Urban, M.J. (1993) *Auditory Subliminal Stimulation: methods*, **Perceptual and Motor Skills 76**: 1103-1106