Building and Maintaining the Bridges of Spoken Language Science

J.Bruce Millar,
Computer Sciences Laboratory,
Research School of Information Sciences and Engineering,
Institute of Advanced Studies,
Australian National University
(bruce.millar@anu.edu.au)

ABSTRACT

The multi-disciplinary nature of spoken language science is often taken for granted. Its analysis, promotion, and maintenance rarely gain centre stage in our minds. It is the thesis of this paper that a conscious focus on the overall shape of our scientific endeavours and their disciplinary ingredients is an important factor in maintaining good progress when exploiting the products, of many scientific traditions. Herein we explore the concepts of disciplines and their interaction. We review the history of the development of understanding of spoken language and the disciplinary foundations that have supported it. We very briefly survey the nature of the supporting disciplines before examining some of the more interesting cross-disciplinary linkages that have been spawned in order to fill out the texture of our current understanding. A brief focus on some key commodities that have been part of cross-disciplinary trade leads into an examination of some bridges that have been negotiated in the author's experience and some of the maintenance issues for the whole field.

INTRODUCTION

Spoken Language, it is said, is the most natural form of communication known to humans. It is true that the vast majority of people can speak and understand speech for the vast majority of their lifetime. It is also true that these people are born with all the hardware required to perform this natural but complex task and need only to tune the software over the early months of life and when confronted with a new language. When we examine what we know about speech, our spoken language science, we find that its description can be highly complex. There is a multiplicity of contributors to this complex description. The number of contributors, and the study disciplines that they represent, have developed through history as the acuity with which we wish to understand the processes of spoken language communication has increased and continues to do so.

Bridges are built to link places that are otherwise isolated, typically, in the physical world, by stretches of water. However, the notion of "building bridges", as evidenced by an internet search on that phrase, has come to refer primarily to the linking of human communities where some form of social, cultural, or generational barrier exists. In the scientific research domain it has come to reflect the linking of our disciplines of study. Our title is therefore predicated on the notion that there do exist barriers within the multidisciplinary domain of spoken language science that they need to be overcome and that we do have disciplinary islands that need adequate bridges to link them effectively.

The process of making connections across disciplines has been graced by the title of the *scholarship of integration* by celebrated educationalist Ernst Boyer (Boyer, 1990). As such it sits alongside his other classifications of the scholarships of *discovery*, *application*, and *teaching*. It is therefore with the theme of integration that I embark on this review.

What is a discipline

A discipline is a set of constraints designed to focus behaviour so that a particular result is made likely. In its many contexts the word itself has connotations of control, order, law, obedience, and punishment.

The specific meaning that we focus on here is "a branch of instruction or learning" or a study discipline. However, it is useful to maintain contact with the characteristics of a study discipline that indeed mirror some of the wider connotations of *discipline* mentioned above. The purpose of defining a discipline of study is to enable an internal regime of order and lawful relationships can be agreed and accepted by its *disciples*. In many cases professional bodies have been established to control admission to the discipline, to control behaviour within it, to reward obedience to its demands, and maybe to punish by expulsion, disciples who disregard its demands. This brief exploration of the nature of an academic discipline goes some way towards revealing the issues that arise when disciplines are linked into a multidisciplinary field.

Disciplinary Island Communities

As our bridging analogy implies, disciplines are, in many respects, like Islands. Their fundamental aim to be self-consistent is often realised via an aim to be self contained and self sufficient. All these qualities are for the good of the discipline, to sharpen its acuity to the features of its target interest. Disciplines develop their own technical language, their own concepts, their own taboos, and their own laxity in areas of little internal concern.

Presented as invited keynote address to International Conference on Speech Processing, Daejeon, Korea, 22-24 August 2001, pp.363-372

What then characterises a so-called *multidisciplinary field*? Is it simply a cluster of islands – an archipelago? If so what is the sea or ocean in which they lie? Does this disciplinary archipelago have a voice that is different to the clamour of island voices speaking out of their independent stances?

In order to gain some insight into these questions, we can look at the natural scene of islands in a sea. I take the example of region of the world in which I live - Oceania, the immediate neighbour of Asia. Oceania is a region of islands stretched across a vast ocean - hence its name. One part of my work is to coordinate interests in spoken language study across the region on behalf of COCOSDA, and through this I have become intrigued by the way that language in the region has developed. Once isolated Island communities, each with their own native language, can communicate, normally, in the first instance, for trade of their local produce and resources. The development of *contact* or *trade* languages has enabled small isolated populations to develop and maintain a level of regional consciousness that leads them beyond simple trading of goods to a sense of loose political cohesion.

Contact languages do not replace native languages that continue to express in great depth the intensity of the life of the island community: All that they hold to be precious to them, their history, their beliefs, their traditions, their relationships. The contact language develops a limited vocabulary and syntax that is adequate for the task of communicating about trade – descriptions of commodities, finance, geography, transport and any other relevant domain.

Such island communities will also normally have an *official* language for the expression of their laws, constitution and treaties in formal terms. This official language may also be used in the later stages of education.

The contact language provides a purposive interface with the outside world creating an infrastructure for coexistence and cooperation. The vocabulary can be sparse compared to *native* or *official* language in which descriptive power and logical precision lie respectively, but it has evolved to fit a need, the creation of a larger functioning unit.

Our multidisciplinary archipelago.

Mapping the linguistic world of a multiplicity of islands onto our technical world of a multiplicity of disciplines, we see many useful parallels. Our official language is the structure of scientific and logical thought expressed with all the power and precision allowed by the natural language of our nation or by the international language of mathematics.

Within each discipline, however, we have a *native* language and mindset, constructed from a history of endeavour, a history of focus, and a history of thinking within a boundary. This native language focuses acutely on those regions of debate that have help shape the discipline so that issues of critical difference are carefully labelled and articulated. In regions where little or no debate has occurred the native language of the discipline may be strangely diffuse.

When a multidisciplinary focus arises in a serious way, we scramble for a *contact vocabulary* that suits our need or at least a glossary of simplified definitions of terms from our *native discipline*. It is notable that there is an increase in the publication of terminology in disciplines such as acoustics with some 2600 entries (Morley, 2000) and indeed forthcoming in our central field of spoken language science with some 10,000 entries (Laver & Asher, forthcoming).

The purpose of this paper is to survey the issues and opportunities surrounding the development of interdisciplinary contact languages driven by the need of spoken language science to be thoroughly interconnected and thoroughly enriched by its multiple disciplinary parts.

HISTORICAL DEVELOPMENT OF OUR DISCIPLINARY ARCHIPELAGO

Today we readily acknowledge that the research field of spoken language communication is multidisciplinary. It is instructive to examine how this has developed over time: Initially from a fascination with the phenomenon of speech as a supremely human artefact, then to the science of the mechanisms and representation of speech, and recently to the analysis, synthesis and manipulation of speech driving a technology to interface humans with machines.

Allen (1953) presents a fascinating insight into the phonetic expertise of ancient Indian descriptive grammarian, Panini. Living about 650 BC, Panini created a description of Sanskrit that has been acclaimed by Bloomfield (1933) as the most complete record of a spoken language, ancient or modern. Whilst the detail of much modern analysis and data recording has since transpired, the stature of the work of Panini in both the linguistics and phonetics of his native language at such an early date is indeed remarkable.

While significant progress of our discipline may have been expected from the burgeoning civilisations of Egypt, Greece, and Rome, it appears that precise and reliable work on spoken language was not a major strength (Clark & Yallop, 1990, p.329). However, ancient Egyptian and Grecian legend implies the concept, and quite probably the implementation, of talking heads on statues that probably worked using suitably hidden speaking tubes (Liénard, 1995). Ohala (2000) records that questions on the basis for a link between sound and meaning, undoubtedly the essence of spoken language communication, and on the nature of

Presented as invited keynote address to International Conference on Speech Processing, Daejeon, Korea, 22-24 August 2001, pp.363-372

the production and perception of speech attracted attention respectively of both philosopher, Plato in 360 BCE, and physician, Galen living between 131-201AD.

In the mid-fifteenth century (1446) Korean scholars at the Palace Library of Classics ("Jip Hun Jun") in Seoul published the first phonetically rational alphabet. This Hangul, or Chosungul, script represents a landmark in the association of phonetics and orthography, which has never been repeated in any significant way. This is perhaps the first realisation of the concept of the "phoneme" whose true definition across the disciplines of phonetics, linguistics and psychology continues into the 21st century.

In the western civilisations the Renaissance was being ushered. The fall of Constantinople in 1453 and the discovery of North America in 1492 heralded the expansion not only of geographical boundaries but also boundaries of knowledge and the desire for understanding of phenomena that were previously uncritically accepted on the basis of traditional philosophy and lack of investigative opportunity.

In the 17th century various studies relating to anatomical and auditory correlates of speech are reported or alluded to in historical records. Bishop John Wilkins, a founder and fellow of the Royal Society, defined the positions taken by vocal organs for each speech sound (Wilkins, 1668). A decade later Samuel Reyher (1679) noted that vowels "not only differ by the shape of mouth and tongue but also by a tone which may be heard when the voice is suppressed and the vowels are produced only by breath" (cited by Kohler, 2000).

In the 18th century we can observe a distinct shift to include the more functional aspects of speaking via the discipline of the *physiologist*. French physiologist C.J. Ferrien published an account of how the vocal folds produce phonation (Ferrien, 1741). Some four decades later C.G .Kratzenstein, a professor of Physics and Medicine, successfully responded to a competition to synthesise five Russian vowels and soon thereafter Wolfgang von Kempelen (1743-1804) demonstrated more extensive acoustic synthesis using an articulatory model.

The 19th century saw the more detailed laying of the foundations of spoken language science. The skills of the *engineer* to fabricate the insights of the physiologist enabled Faber's 'Euphonia' to synthesise with greater sophistication than von Kempelen. However it was in the realm of *acoustics* that more strides were taken to build greater understanding. Willis (1829) described the first systematic acoustic investigation into the nature of vowels and made the critical link between the quality of the vowel and the resonances of the mouth cavity that produced them. Wheatstone (1837) recognised the fact of the "multiple resonance" of a single but complex resonant cavity such as the vocal tract, while Helmholtz (1885) linked higher pitched resonances with the constrictions within a major cavity. However, it was a *physiologist* (Lloyd, 1891) who pointed out "the probability that every cardinal vowel derives one chief resonance from the anterior or oral part of its articulation, and another from the posterior or pharyngeal part". Lloyd studied the form of the sound waves recorded phonographically and on measuring the resonant frequencies deduced that the identity of the vowel was based not on the absolute pitch of the two major resonances but on their mutual interval.

The interaction of *physiology*, training in elocution, orthographic innovation, *acoustics*, and eventually the *engineering* of the telephone all within the family of Alexander Melville Bell (1819-1905) characterises the 19th century multidisciplinary contribution to spoken language science (Bell, 1867; 1876; 1879). A highly significant heritage of the 19th century, informed by advances in *acoustic* and *physiological* understanding was the realisation of speech sounds in electrical (Bell, 1876) and then visible form (Lloyd, 1891). A further 19th century inter-disciplinary dimension was realised by Rosapelly (1876), a *linguist* conducting *physiological* measurement on production of nonsense words to examine both "sound change" and "communication disorders" (cited by Ohala (2000)).

In the 20th century our science was carried forward by the growing sophistication of *acoustic* understanding, the *engineering* based on the electrical and visual representation of the acoustics. The disciplines of *anatomy* and *physiology* receded in prominence, and by mid-century *electrical* analogues of the *acoustic* patterns took centre stage (Dudley, 1937) followed rapidly by graphical and parametric control (Lawrence, 1953; Cooper, 1953) inspired by the *engineering* of graphical displays of these acoustic patterns (Potter et al, 1946). This was quickly followed by the higher order information modelling of speech processes (Holmes et al, 1964) afforded by computer technology that allowed the acquisition, management and analysis of the acoustic stream of speech. It was in the 1970s that early speech recognition projects began to proliferated in the newly emerging domain of computer science. This continued in the 1980s with the emergence of a dominant *information science and engineering* technology of hidden Markov models. The 1990s saw the emergence of speech data resources on massive scale to satisfy the appetite of stochastic models of written and spoken language used within automatic speech recognition systems. The development of these data resources has itself become a significant area of cross discipline trading. Consortia and conferences have developed to link the *linguistic, acoustic, phonetic, engineering*, and *information science* expertise required for good design of these resources.

The ancient world looked at spoken language through the lenses of *philosophy* and *anatomy*, then starting at the Renaissance, when cultural and religious freedoms started to blossom, notions of spoken language were

Presented as invited keynote address to International Conference on Speech Processing, Daejeon, Korea, 22-24 August 2001, pp.363-372

expressed in the more functional domains of articulation, *physiology*, and *acoustics*. By the second half of the 20th century the focus was securely in the realms of *acoustics*, *phonetics* and *linguistics*: the domains expressing the physical realisation of the transmitted signal on the one hand, and the structures within which these signals could represent meaning. Towards the 21st century the influence of *information sciences* and *engineering* were starting to be felt strongly, and my guess is that the influence of the *social sciences* will soon be just as evident, as will a re-examination of the disciplinary interaction that has landed the technology, built upon some simplified products of many disciplines, on to what appears to be a performance plateau.

THE MAJOR ISLANDS AND ISLAND GROUPINGS IN OUR ARCHIPELAGO

The history of disciplinary focus on spoken language has resulted in some nine major disciplines having a significant perspective on the science of spoken language. These disciplines have become established on independent grounds, and represent the disciplinary islands of our multidisciplinary archipelago. Some of these disciplines have developed as small clusters of islands, or island groups, that have many self-consistent concepts and paradigms, and yet relate individually with other islands outside the group. An example of this would be the islands of *acoustic phonetics*, *articulatory phonetics*, and *auditory phonetics*, all describing the sounds of speech but with the flavours of *acoustics*, *physiology*, and *psychology*. We will now examine the core disciplines that have arisen in the history of our field.

Phonetics

Phonetics is the science of speech sounds. It has three major dimensions which may be characterised the *contemporary*, the *diachronic* and the *synchronic.* The contemporary focuses on the "here and now" analysis of the sounds of speech in processes of speech communication. The diachronic focuses on the changes in the sounds of speech, both historically and in currently active processes. The synchronic focuses on the diversity of the sounds of speech spread across all phonetically analysed languages, giving insight into how different languages will pose different issues for the processing of its spoken form.

A critical tool of phonetics is *phonetic transcription* in which unambiguous symbols are assigned to all the sounds of speech. Phonetic labels are essentially language-independent as they focus on the sounds and not the structure of spoken language. Phonemic, or broad phonetic labels, are language dependent as they subsume phonetic detail that is not distinctive within the language.

The study of phonetics has a long history. Its major instruments over most of its history have been the trained ear and the trained vocal tract. The auditory analysis and the synthesis of vocal performance has been an essentially unique contribution of the phonetician to the study of speech. In recent times "impressionistic" phonetics, based on the use of the trained ear and trained vocal tract, has been strongly enhanced by "instrumental" phonetics. The use of instruments to measure the sounds impinging on the ear, air-flow in the vocal tract, electrical activity in vocalic musculature, and the use of x-rays, electrical current, or light sources, judiciously directed or located in the vocal articulation environment, has given greater explanatory power to the products of the phonetician.

Linguistics

The study of the components and structure of spoken languages has generated a wide range of schools of thought and hence systems of description. For the purpose of our survey however, we can characterise linguistics as the science of semiotic structure of spoken language. This structure can indeed be subdivided into linguistics speciality areas such as lexicography, syntactic analysis, discourse analysis and semantics.

It can also be significant that there are many traditions or "schools of thought" with the linguist discipline. To the extent that these variants use different conceptual frameworks of reality, they could perhaps be considered as parallel disciplines with the same name. Such internal disciplinary divergences can often be overlooked in the necessarily simplified external view of a discipline that seeks to trade ideas across disciplinary boundaries.

Acoustics

As one of the three primary sensual domains of classical physics (heat, light, and sound), and the domain in which speech is physically transmitted from its source to its target, acoustics can perhaps claim a fundamental place in the disciplines of speech. However beyond the physics of generation and transduction of acoustic vibrations, there lies analysis of the signal within the constraints of acoustic space as the wave patterns of our signal interact with those of competing signals and the varying acoustic characteristics of the medium through which they move including objects through and around which they move. The acoustic scene of speech transmission contains elements that can often be ignored but as speech transmission in adverse environments comes into focus the full complexity may be usefully exploited.

Audiology

The study of hearing sits somewhat intermediately between the disciplines of acoustics and psychology in our multidisciplinary space, and may boast the longer title of auditory psychophysics. The contribution of audiology to the disciplines of speech occurs at several levels. A most obvious one is the taking account of individual differences in defining the characteristics described as "normal hearing", a term used frequently and maybe rather loosely in a large body of speech perception literature. Detailed understanding of the sensitivity of hearing at different frequencies, the

Presented as invited keynote address to International Conference on Speech Processing, Daejeon, Korea, 22-24 August 2001, pp.363-372

masking of frequency sensitivity by concomitant sounds at different frequency or at previous times, the assessment and scaling of the psychophysical variables of loudness, pitch and timbre, all enrich the transform between the discipline of acoustics and that of perceptual phonetics.

Psychology
Having listed the disciplines of audiology and phonetics, both of which interact strongly with the psychology of perception, we now view psychology in its cognitive domain. Here we find the dimensions of memory, representations of linguistic and phonetic knowledge, processes of learning and decision making, the impact of reasoning, of emotions and of interactions with others. As the range of environments in which we seek to understand our spoken language science increase, many of these factors and their interactions with related disciplines come into focus.

Physiology
The discipline of physiology is here taken to subsume anatomy for our purpose as it focuses on the functional roles of bones, muscles, cartilage and soft tissue in the production of speech. It logically also includes the physiology of hearing, but that we subsume under the science of hearing – audiology. It is in the physiology of speech production that we gain perspective on the reasonable constraints on speech acoustics. It is the physical processes of aerodynamics, muscle tension, articulator acceleration and deceleration, and the pressure resource of the lungs that impact on how the acoustics of speech actually work in an individual speaker. Given the complexity of the human vocal apparatus there is more than one way that a given acoustic result may be achieved. Matters of individual habitual practice or innovative experiment are most directly represented in the physiological domain.

Sociology
The sociology of spoken language can lay claim to the fundamental status of its raison d'être. Yet it is only as we move from the more formal styles of spoken language communication, that are adequately encoded using the more fundamental paradigms of *phonetics* and *linguistics*, that this discipline has its impact. This progression in complexity of spoken language description may be observed in the development of speech technology in recent decades as it has grown through the recognition genres of simple commands and isolated words, the natural dictation of text, to interactive conversation using spontaneous speech. It is at this latter level where the person-to-person awareness in terms of familiarity, status, personality, and attitudes start to impact on the form of language and the tone of voice used.

Information Science
The modelling of information transmission via spoken language between humans has been studied for a long while within the sub-domain of psychology labelled human information processing, however the generic issue of speech information processing, including transformation of this information into novel domains such as the artificial neural domain of the bionic ear or the acoustic vector domain of the automatic speech recogniser, has created a speciality focus for the information scientist. The transformation of the acoustic stream of speech into a host of different representations, each having different mathematical properties and therefore capable of wide exploitation in the quest for equivalent speech information processing to the human system, creates an information processing discipline that is uniquely different from its disciplinary neighbours.

Engineering
The implementation of artificial systems of spoken language processing has grown strongly in recent years, yet the residents of this disciplinary island have been active with their *physiologist, acoustician* and *information scientist* neighbours for most of the history of scientific examination of speech. The measurement of the production of speech and artificial synthesis techniques have almost exclusively required the insight and expertise of the engineer to implement them. Increasingly the work of engineers is moving from the domain of the physical world into the domain of *information engineering* in which a wide range of general theoretical models, such as the now well-known hidden Markov modelling, are implemented within the sub-speciality of *software engineering*.

TRADING IN THE MULTIDISCIPLINARY OCEAN

In addition to the "landmass" or "island" analogy with established professional structures, we should also carry in our minds other models such as that of "trading" as used by Ohala (2000) in relation to the development of phonetics as a discipline. The ocean in which our disciplinary islands lie is a medium for trade, for the exchange of ideas in which broad questions may be asked and in which partial answers may be bought and sold. As in the physical world of trade there exist both many incentives and many barriers to the equitable exchange of benefits.

We can therefore find, in addition to our major island groupings that may comprise a cluster of minor islands, some distinctive regions in which reside investigators who trade intensively in their immediate region of major island groups. These concentrated patterns of interdisciplinary linkage have developed because they have proved to be especially fruitful. Quite often this has been found where two major disciplines have found either multiple points at which they can interact or else multiple applications to which they can both be applied.

Laver (2000) identifies the linkages that have been particularly significant, and indeed reciprocal, for the discipline of phonetics with acoustics, psychology,

Presented as invited keynote address to International Conference on Speech Processing, Daejeon, Korea, 22-24 August 2001, pp.363-372

	Audiology	Physiology	Phonetics	Linguistics	Sociology	Acoustics	Psychology	Engineering	Info.Sciences
Hearing	medical & surgical	monitor deficits	monitor linguistic deficits	monitor social deficits	hearing aid design	multimodal cue trading	microphones, hearing aids, bionic ear	info transmission analysis
Physiology	**Biological mechanism**		cerebral hemisphere	derive population norms	cochlear function		electrical stimulation	articulatory modeling
Phonetics		**Sounds of Speech**	phonology	accents, prestige forms, speaker characteristics	identifying phonetic objects	checking perceptual boundaries		modeling speech sounds
Linguistics			**Semiotic structure of speech**	refining models of communication	long-term character	lexical access etcetera		language modelling
Sociology				**Social Role of Speech**	speaker space analysis	dialogue modelling	Speech Technology acceptance	
Acoustics					**Analysing Sounds**	basic auditory scenes	sound acquisition	data resources & structures
Psychology						**Perception & Cognition**	System useability studies	HCI using speech
Engineering							**Creating Speech Systems**	software engineering
Info.Sciences								**Information modelling of speech**

TABLE 1. Disciplinary interactions relating to spoken language science**.**

sociology, and linguistics. Reciprocity is, of course, a strong factor in linking. It implies a bridge carryingsignificant two-way traffic. It can be observed when each discipline gains a foothold in each others territory, so that for instance the role of phonetics in describing the signalling of "turn taking" in a conversation is acknowledged in sociology, and the role of sociology in describing the concept of turn taking and the characterisation of social factors that influence the way people talk when in conversation is acknowledged in phonetics. In each case added structure is supplied to the description of the conversational scene.

Some further examples

One such intermediate island is the island of *phonology*. Phonologists trade intensively within the regions of *phonetics* and *linguistics*. Their products resemble linguistically structured systems of phonetic aggregation. Generative phonology can create a non-physical "mentalist" view of the phoneme – the physical reality of sound is created from the application of phonological rules to an underlying abstract model (see Chomsky and Halle 1968 p14).

The phonological system of a language creates more than structure. Together with the lexicon of the language it defines the degree of phonetic variation between or within speakers that can be tolerated as acceptable. The phonological system can therefore create phonetic space in some areas of human productive repertoire allowing freedom for individual expression without interfering with linguistic category assignment, whereas in other areas phonetic space can be very limited thus requiring high precision to avoid ambiguity.

Another intermediate island is the island of *psycholinguistics*. Psycholinguists trade intensively within the regions of *psychology* and *linguistics*. Their produce helps refine linguistic study by drawing conclusions about some of the underlying processes that seem active in the generation and reception of linguistic stream (Cutler, 1996). One example of such processes is the set of ways that language perceivers break up the acoustic stream into words. Such understanding is essentially language independent. Only certain features of this kind of understanding will apply to a specific

Presented as invited keynote address to International Conference on Speech Processing, Daejeon, Korea, 22-24 August 2001, pp.363-372

language but evidence from many like and unlike languages are involved in deriving this cross-language perspective.

There is indeed also the very relevant but often overlooked area of *psycho-phonetics*. This is the use of both *acoustic-phonetic* and *psycho-acoustic* expertise to explore the detailed dependencies for perception of speech sounds of the various features of the speech acoustic stream. This area was a focus of work at the Pavlov Institute of Physiology in St Petersburg (e.g. Chistovich & Lublinskaya (1979)). Whereas the psycho-acoustician has mapped the perceptual response to sound in general, it is very significant that another comparable level of disciplinary interaction can reveal important relationships between speech-related acoustic primitives such as vocal-tract resonances. Chistovich and Lublinskaya showed that the perception of two vocal tract resonances was equivalent to that of one vocal tract resonance placed at the spectral centre of gravity of the two resonances on the condition that the two resonances were within 3.5 bark of each other. This was established in the region where the first and second formants of a vowel can be close to one another. The most plausible physiological mechanism advanced was spatial integration over wide intervals along the cochlea scale.

Returning to the production of speech, there is broad disciplinary content in the study of how *information* is laid down in the *acoustic* signal. There has been an ongoing multidisciplinary exercise of understanding the mechanisms. At the most basic level we have the bio-mechanical level which is controlled by the masses and muscular forces of the *anatomy and physiology* of the vocal apparatus which dictate limits as to how and how rapidly its configuration can change. Certain articulators are less well equipped for rapid movement and thus in the ongoing stream of speech they lag behind other articulators, on account of their local characteristics. This generates a time smearing of the acoustic characteristics for which they are responsible. Similarly for optimum signalling of those phonetics elements for which they are responsible such articulators need to be actuated prior to the time when their presence will be crucial. This anticipatory coarticulation can only be understood directly by examining the timing of neuro-muscular commands to the responsible muscles. It is instructive here to note that the most visible "export" of phonetics, the phonetic symbol, is based on the "convenient fiction" of the phonetic "segment", according to Laver (1994). However, in the greater, but largely hidden, interior of the discipline there lie the refinements of non-linear and auto-segmental approaches to the description of the phonetics stream. Both these internal models and the external *physiological phonetics* represent a highly complex area that is poorly understood as we enter the 21st century. Adding further complexity, Clark and Yallop (1990) indicate that there can often be *sociological* factors, such as prestige forms produced in deference to a perceived authority, that will cause the overriding of flexibilities allowed by either *phonological* or *physiological* factors.

Such areas of conceptual trade between disciplines could be expanded upon much further, but the above examples serve to illustrate some of the facts and principles involved. The study of the phenomena of the world in which we live according to scientific principles is indeed a continuum. However, human understanding of these phenomena has been gradual as concept has been added to concept, discipline to discipline, as useful links have been made. Our own multidisciplinary area illustrates this process very well both in its history and in its ongoing development.

We will now focus on some specific disciplinary bridges that have been important to my own work and on some mechanisms that have already contributed to good bridge maintenance within our currently vibrant science and technology.

SOME SPECIFIC BRIDGING CONCEPTS

The phoneme.

The phoneme is one of the most shared concepts across the spectrum of the speech sciences and yet also one that is often misunderstood. The concept appears to have arisen first with Hindu grammarians working with Sanskrit in the 5th century BC, then clearly articulated in the work of an unknown Icelandic grammarian working in the 12th century AD, effectively implemented by Korean scholars in the Hangul alphabet in the 15th century AD, but not given its current name until used by European linguists in the late 19th century – albeit with various shades of meaning (Clark & Yallop, 1990). Today it is used very widely to refer to the atomic unit of the phonology of the language. Hence it is essentially a functional unit of *linguistic* origin, that is, however, capable of multiple kinds of phonetic realisations. It is expressed stochastically by *engineers and information scientists* as a mixture of Gaussian distributions used to describe the variable acoustic vectors derived from processing the acoustic signal of large numbers of phonemic realisations in running speech.

The fact that the phoneme does not have a one-to-one relationship with the *acoustics* and *phonetics* in which it is realised, and the fact that phonemic information is encoded in *acoustics* in a fashion that has its basis in *anatomy, physiology, neurology, linguistics*, and *sociology* can easily be overlooked by the engineer who finds that a stochastic model of the phoneme (albeit a tri-phone rather than a mono-phone model) can sustain an adequate system for the automatic recognition of continuous speech. The notion in phonetics of the constituents of a phoneme standing in complementary distribution (ie occurring in mutually exclusive contexts) is matched only to a certain degree by the triphone model that normally has its explanatory base in the realms of coarticulation. Clearly effects that are

conditioned on non-adjacent phonetic segments are not accounted for by this limited context model.

The formant

The formant is another good example of a concept that has often been traded outside the discipline of *acoustic phonetics* in which it came to prominence. In this sub-discipline it comprises concepts of order within the set of spectral features evident on a spectrogram (or perhaps more historically - of order within what can be heard distinctly as an auditory feature in parallel with others). It reflects a component of the resonance qualities arising in the oral tract and perceived in the make-up of vowel quality that develops across a vocalic syllable. It has been a very attractive "commodity for trading" owing to its extreme simplicity for the description of both vowels and the vocalic nuclei of syllables. However, when the *engineer or information scientist* has tried to harness this "import" for their own benefit, they have discovered that they are working with a rather incomplete understanding of the originating discipline. There maybe other spectral maxima observed in the acoustic description of speech but the formant has that more purposive role of signalling the vocalic quality of the syllable. The complexity difference between a relatively simple spectral peak picker and a true formant tracker illustrates the conceptual difference.

SOME PERSONAL BRIDGES

A few examples show how in the experience of the author careful relating across disciplines has enriched his level of understanding of the total phenomenon that we study.

Acoustics to Psychology – Vowel Perception

The psychology of perception provides a basic set of relationships between acoustic variables and psychological variables (eg frequency and pitch, intensity and loudness). It then shows us that these basic psycho-acoustic variables have more complex relationships with their acoustic counterparts (eg the secondary dependence of loudness on frequency, masking effects of sounds in that are adjacent time or frequency). Then when we consider the components of speech sounds within a specific language framework, we have another level of complexity – that of categorical perception. When I present a complex spectrum that represents a set of resonances excited by a broadband excitation signal (all of some appropriate frequencies), the hearer can be persuaded that they are listening to the output of an artificial vocal tract, s/he recognises a vowel sound. Which vowel sound is heard will depend on several factors: the assumed size of the vocal tract, the assumed accent or phonological variety spoken by the vocal tract, the position of the acoustic stimulus within the space available for a particular vowel based on those assumptions. With artificial stimuli it is still possible that there is uncertainty about the identity of the intended sound. Interestingly we then find that the amount of context provided within the stimulus to bias the assumptions of the listener towards one alternative or another can play a significant role. In 1972 we published results showing how artificial vowel stimuli were identified with increased clarity when placed within an artificial syllabic frame (Millar & Ainsworth, 1972). It was claimed that this frame gave normalising information that enabled the speech perceptual system to work efficiently. This experience provided evidence that a simple speech perception experiments, although an essential starting point, can ignore important cues from the wider sphere of spoken language science.

Acoustics to Neurophysiology or Psychophysics – Early Cochlear Implant experiences

An early lesson in issues of disciplinary transfer was learned when I first consulted on the issues relating to the development of a direct acoustic-neural hearing prosthesis. Here, on the one hand we had the acoustic description of speech and on the other we had the physiological description of information transfer in the organ of hearing, the cochlea. There existed a small vocabulary "trade language" of concepts that involved descriptions of the acoustic frequencies judged by audiologists to be important for speech perception, and also the mechanisms involved in neural excitation in the cochlea together with an understanding of the tonotopic organization of the cochlea. In this "language" the cross-disciplinary procedure was described in terms of filtering the signal into bandlimited channels, generating a stochastic pattern of pulses for each channel related to the acoustic energy in that channel, and the application of these pulses with careful attention to charge balancing and amplitude. It quickly became apparent that this language was inadequate in some way. The experimental variables were not capable of controlling the perceptual impact of electrical stimulation delivered in this way, even though all aspects of biological safety were well controlled.

The way that this impasse was bypassed was by building an alternative bridge using a "trade language" of psycho-physics. Within this disciplinary region the islands of physics and psychology had established a language of trade that involved the detection and magnitude scaling of physical variables as experienced by the human sensory system. An extension of this trade language enabled us to look at a common acoustic phenomenon, speech signals, through two *psycho-physical* lenses. One lens was the psycho-physics of electrical stimulation of neural tissue in the cochlear, such that the scaling of perceptual experience with respect to the intensity of pulsatile stimulation, to variation of the site of that stimulation, and to the pulse rate of that stimulation was determined. The other lens was the psycho-physics of natural stimulation of acoustic origin in which the perceptual experience of acoustic intensity as loudness, and of acoustic frequency as pitch, with the latter subdivided into fundamental tonal pitch on a scale of

Presented as invited keynote address to International Conference on Speech Processing, Daejeon, Korea, 22-24 August 2001, pp.363-372

low to high and spectral shape timbre pitch on a scale of dull to sharp. By mapping the specific psycho-physical characteristics of an implantee using the first lens onto the characteristics of normal hearers using the second lens, it was possible to achieve a viable speech coding scheme (Millar et al., 1990; 1992). This involved some very early *psycho-phonetic* results surrounding the concept of the single equivalent formant, but this work pre-dated the more refined work of Chistovich and Lublinskaya (1979).

In retrospect it can be seen that an alternative trade route between disciplines using alternate trading languages was the mechanism that was effective in this situation. The original route may still be valid but clearly required a much richer trading language to achieve understanding between the *engineer* and the *physiologist*.

Comprehensive Spoken Language Description

Literature on automatic speech recognition has been characterised over most of the past half century by a repeated theme – that there are aspects of speech that we have not accounted for but which are most likely used by the human system whose performance we attempt to emulate. Early on it was higher order linguist effects, latterly it has been aspects of spontaneous speech, and all through have been uncertainty about the handling of an adequate range of individual speaker characteristics.

I have been concerned over the last decade or so with the cross-disciplinary description of spoken language data (e.g. Millar, 1989; 1992; 1998). Each discipline has its own perspective on the sources and patterns of variance that are found in spoken language data. These range from the awareness of background noise and a speakers' response to it, the assumptions about the listeners prior knowledge, the pervasiveness of articulatory settings, the impact of abnormal hearing, the impact of abnormal anatomy or physiology, the relationship of the speaker and the listener, the impact of the physical size of the vocal apparatus, and the impact of the state of tension of vocal musculature. The cumulative effect of all these matters may well still be beyond our analytical ability to represent in a useful way, but the collection of very basic but broad information to attach to archived data opens the way for such data to be used within a range of disciplinary frameworks. A most common rejection of re-use of data is based on the fact that, in the disciplinary view of the prospective user, inadequate controls were applied in its collection. Such a lack can be mitigated to some degree by descriptions that make explicit what degree of variance may be expected.

MAINTAINING CROSS-DISCIPLINARY BRIDGES - ASSOCIATION OF DISCIPLINES

When I first entered speech research in 1964, it did not have a clear articulation of its multidisciplinary base. My first realisation in practice of this base was by observing how work in this field attached itself to the deliberations of adjacent disciplines. If I had entered via the discipline of phonetics then I would have perhaps had a more mono-disciplinary entry as the International Congress of Phonetic Sciences was indeed firmly established but in those days moving only slowly towards the instrumental experimental phonetics that was the forerunner of much modern speech technology. The multidisciplinary flavour was promulgated however by some visionary laboratories, such the Haskins Laboratories in New Haven, USA.

In my experience in Australia since 1970, the speech research community has moved forward in significant ways every few years refining its disciplinary mix. Through most of the 1970s a few academics, drawn mostly from linguistics, phonetics, and computing, and fewer scientists working in the software industry on interactive techniques, maintained a loose linkage through an occasional newsletter. In 1978, an informal Australian Speech Research Association was formed. The number of people involved increased and the disciplines expanded to include some people in psychology and in clinical areas. At this stage the engineering component of our concerns was largely found in the instrumentation available in research laboratories, hence the latest hardware and software for analysing, synthesising, displaying and managing speech data were discussed with vigour.

In 1984, a determined effort was made to define a broad-based national focus that resulted in the launching of our first multidisciplinary national conference in the now well-established speech science and technology (SST) series (1986-). This was followed in 1988 by the formation of an incorporated Australian Speech Science and Technology Association (ASSTA) with a membership crossing many disciplinary boundaries in speech science and with a significant interest in speech technology.

In the wider world the development and coverage of journals and then conferences servicing the field give testimony to a century of disciplinary expansion. The centrality of acoustics strongly represented by the Journal of the Acoustical Society of America (1929-) and the International Commission on Acoustics (1951-), and phonetics represented by the International Phonetic Association (1886-) and the International Congress of Phonetic Sciences (1932-) are clear. In the second half of the 20th century we have seen the formation of the IEEE covering engineering aspects of spoken language and the emergence of multi-disciplinary journals and conferences (Speech Communication (1982-), Computer Speech and Language (1987-), Eurospeech (1989-), and the International Conference on Spoken Language Processing (1990-)).

In 1997, the Personal Computer world was hit by offerings of low-priced natural speech recognition software, and a new dimension of interest in our field –

Presented as invited keynote address to International Conference on Speech Processing, Daejeon, Korea, 22-24 August 2001, pp.363-372

the user of speech technology – was born. In 2001, there are forums and conferences being held in Australia at which the business world is debating the pros and cons of introducing telephone-based speech technology.

Over these last 30 years our field has moved from the research interests of a small group of academics who were prepared to embrace multidisciplinary projects, to a technology that still has rough edges but which is being taken seriously by the business sector of the economy. Does this herald the need for our multidisciplinary archipelago to have some of its bridges upgraded. Perhaps to link more effectively to the disciplines of technology-use. The legal and sociological issues of the use of our imperfect but increasingly useful technology are starting to impact on the business world. After centuries of focus on written language as its authoritative form, we are on the brink of a revolution where its far more expressive spoken form is now capable of capture, transmission, and validation. Spoken language science needs to strengthen its disciplinary bridges for these challenges to be met in a scientifically principled manner.

REFERENCES

Allen,W,S. (1953) *Phonetics in ancient India*, London: OUP

Bell,A.M. (1867) *Visible Speech: the science of universal alphabetics.*

Bell,A,G. (1876) *Alexander Graham Bell's Telephone Patent Drawing and Oath*, Patent No.174465.

Bell,A.G. (1879) *Vowel Theories,* American Journal of Otology, Vol.1.

Bloomfield (1933) *Language***,** Holt: New York.

Boyer,E. (1990) *Scholarship revisited: the priorities of the professoriate*, The Carnegie Foundation for the advancement of teaching, New Jersey.

Cooper,F.S. (1953) *Some Instrumental Aids to Research on Speech*, Report on the Fourth Annual Round Table Meeting on Linguistics and Language Teaching, Georgetown University Press, pp. 46-53.

Chistovich, L.A., Lublinskaya, V.V. (1979) *The 'Center of Gravity' effect in vowel spectra and critical distance between the formants: psychoacoustical study of the perception of vowel-like stimuli*, Hearing Research, Vol.1, pp.185-195

Chomsky,N., Halle,M., (1968) *The Sound Pattern of English*, Harper Row: New York.

Clark,J.E., Yallop,C (1990) *An introduction to phonetics and phonology*, Oxford:Blackwell.

Cutler,A. (1996) *The comparative study of spoken language processing*, Proc. ICSLP'96, Vol.1, p.1.

Dudley,H. (1939) *The Vocoder*, Bell Labs. Record, Vol.17, pp.122-126.

Ferrien,C.J. (1741) Mem. Acad. Paris, Nov 15, pp.405-432.

Helmholtz, H. (1885) *On the Sensations of Tone as a Physiological Basis for the Theory of Music*, (reprinted as *On the Sensations of Tone*, tr. Alexander Ellis, Dover: New York, 1954).

Holmes,J.N., Mattingley,I., Shearme,J. (1964) *Speech synthesis by rule*, Language & Speech, Vol.7, p127-143.

von Kempelen,W. (1791) *Le méchanisme de la parole, suivi d'une description de la machine parlante*, Vienne.

Kohler,K.J. (2000) *The future of phonetics***,** Journal of the International Phonetic Association, Vol.30, pp.1-24.

Laver,J. (1994) *Principles of phonetics*, Cambridge: Cambridge University Press.

Laver,J., Asher,R.E. (forthcoming) *Encyclopedic Dictionary of Speech*, Cambridge,MA: Blackwell.

Lawrence,W. (1953) *Synthesis of speech from signals which have a low information rate*, In "Communication Theory", (ed. W.Jackson) Butterworths:London, p460.

Liénard,J-S. (1995) *From speaking machines to speech synthesis*, Proc. 12th Int. Congress of Phonetic Sciences, Aix-en-Provence, Vol.1, pp.18-27.

Lloyd, R.J. (1891) *Genesis of vowels*, Journal of Anatomy and Physiology, Vol.31, p249.

Morley,C. (2000) *Dictionary of Acoustics*, Academic Press

Millar,J.B., Ainsworth,W.A. (1972) *Identification of Synthetic Isolated Vowels and Vowels in h-d Context,* Acustica Vol.27, pp.278-282

Millar,J.B. (1989) *Design and use of a national speech database,* In "Proceedings of the ESCA workshop on Speech Input/Output Assessment and Speech Databases", Noordwijkerhout, 20-23 September, pp.2.5.1-2.5.4.

Millar,J.B., Blamey,P.J., Tong,Y.C., Patrick,J.F., Clark,G.M. (1990) *Speech Perception,* In ``Cochlear Prostheses", Edited by G.M.Clark, Y.C.Tong, J.F.Patrick. Churchill Livingstone:London.

Millar,J.B., Blamey,P.J., Clark,G.M., Dowell,R.C., Patrick,J.F., Seligman,P.M., and Tong,Y.C. (1992) *Speech processing for cochlear implants*, In "Advances in Speech Hearing and Language Processing", Volume 2, pp.217-251. JAI Press: London.

Millar,J.B. (1992) *The description of spoken language*, In "Proceedings of 4th Australian International Conference on Speech Science and Technology" Brisbane, Australia, pp.80-85.

Millar,J.B. (1998) *A structure for comprehensive spoken language description,* In "Proceedings of First International Conference on Language Resources and Evaluation (ICLRE'98)", Granada, Vol.2, pp.1303-1308.

Ohala,J.J. (2000) *Phonetics in the free market of scientific ideas and results*, J. International Phonetic Association, Vol.30, pp.25-29.

Potter,R.K., Kopp,G.A., Kopp,H.G. (1966) *Visible Speech*, Dover: New York. (first published in 1946)

Reyher,S. (1679) *Mathesis Mosaica, sive Loca Pentateuchi Mathematica Mathematice Explicata*.

Wheatstone,C. (1837) Westminster Review, Vol.28, p.37.

Wilkins, J. (1668) *An essay towards the real character and a philosophical language*, Royal Society: London.

Willis, R. (1829) Transactions of the Cambridge Philosophical Society, Vol.3, part 1, x, p.231.

Presented as invited keynote address to International Conference on Speech Processing, Daejeon, Korea, 22-24 August 2001, pp.363-372

Presented as invited keynote address to International Conference on Speech Processing, Daejeon, Korea, 22-24 August 2001, pp.363-372